STATISTICAL GUARANTEE FOR NON-CONVEX OPTIMIZATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Botao Hao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Guang Cheng, Chair

Department of Statistics, Purdue University

Dr. Hyonho Chun

Department of Mathematical Sciences, KAIST

Dr. Jean Honorio

Department of Computer Science, Purdue University

Dr. Qifan Song

Department of Statistics, Purdue University

Approved by:

Dr. Jun Xie

Head of the Graduate Program, Purdue University

To my family.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisor Prof. Guang Cheng for his support of my Ph.D. study such that I have enough time to purely focus on research for my five and half years. I really learnt a lot, independence, critical thinking, rigorous attitude for research and scientific writing. I deeply appreciate the guidance I have received from professors in Department of Statistics at Purdue University. Especially, I would like to thank my thesis committee members: Professor Hyonho Chun, Professor Jean Honorio and Professor Qifan Song, for their insightful comments and questions that help me complete my thesis. Special thanks to Professor Hao Zhang for his numerous supports on my academic travels. I would also like to acknowledge group members in the Big Data Theory Group, including, Dr. Zhuqing Yu, Dr. Shih-Kang Chao, Dr. Meimei Liu, Dr. Ching-Wei Cheng, Yang Yu and others, for many valuable discussions on research problems over the past five years. I also deeply appreciate generous help and company from friends at Purdue. Special thanks to Dr. Wei Sun. Without his help, I couldn't have my first paper published and first internship offer. I would also like to thank my roommates: Dr. Min Ren and Dr. Yun Lu, who offer me unlimited help and a lot of funs during my years at Purdue. In the end, I thank my family for their selfless support of my research.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Hao, Botao Ph.D., Purdue University, December 2019. Statistical Guarantee for Non-convex Optimization. Major Professor: Guang Cheng.

The aim of this thesis is to systematically study the statistical guarantee for two representative non-convex optimization problems arsing in the statistics community. The first one is the high-dimensional Gaussian mixture model, which is motivated by the estimation of multiple graphical models arising from heterogeneous observations. The second one is the low-rank tensor estimation model, which is motivated by high-dimensional interaction model. Both optimal statistical rates and numerical comparisons are studied in depth.

In the first part of my thesis, we consider joint estimation of multiple graphical models arising from heterogeneous and high-dimensional observations. Unlike most previous approaches which assume that the cluster structure is given in advance, an appealing feature of our method is to learn cluster structure while estimating heterogeneous graphical models. This is achieved via a high dimensional version of Expectation Conditional Maximization (ECM) algorithm (1). A joint graphical lasso penalty is imposed on the conditional maximization step to extract both homogeneity and heterogeneity components across all clusters. Our algorithm is computationally efficient due to fast sparse learning routines and can be implemented without unsupervised learning knowledge. The superior performance of our method is demonstrated by extensive experiments and its application to a Glioblastoma cancer dataset reveals some new insights in understanding the Glioblastoma cancer. In theory, a non-asymptotic error bound is established for the output directly from our high dimensional ECM algorithm, and it consists of two quantities: statistical error (statistical accuracy)

and optimization error (computational complexity). Such a result gives a theoretical guideline in terminating our ECM iterations.

In the second part of my thesis, we propose a general framework for sparse and low-rank tensor estimation from cubic sketchings. A two-stage non-convex implementation is developed based on sparse tensor decomposition and thresholded gradient descent, which ensures exact recovery in the noiseless case and stable recovery in the noisy case with high probability. The non-asymptotic analysis sheds light on an interplay between optimization error and statistical error. The proposed procedure is shown to be rate-optimal under certain conditions. As a technical by-product, novel high-order concentration inequalities are derived for studying high-moment sub-Gaussian tensors. An interesting tensor formulation illustrates the potential application to high-order interaction pursuit in high-dimensional linear regression.

1. INTRODUCTION

The integration and interdiscipline between statistics and optimization are urgent and productive. More and more statistical tools are being developed by borrowing strengths from optimization, while optimization is looking to statistics for new insights, speed, and robustness. In the era of big data, traditional statistics meets the curse of heterogeneity and the emergence of new data types, such as matrix-value or tensor-value data. In the optimization community, non-convex optimization has been shown to handle these challenges efficiently but lacks a statistical guarantee. In this thesis, we develop some new optimization tools for two classical non-convex statistical problems as follows and provide the solid theoretical guarantee.

1.1 Simultaneous Clustering and Estimation of Heterogeneous Graphical Models

Graphical models have been widely employed to represent conditional dependence relationships among a set of variables. The structure recovery of an undirected Gaussian graph is known to be equivalent to recovering the support of its corresponding precision matrix (2). In the situation where data dimension is comparable to or much larger than the sample size, the penalized likelihood method is proven to be an effective way to learn the structure of graphical models (3; 4; 5). When observations come from several distinct subpopulations, a naive way is to estimate each graphical model separately. However, separate estimation ignores the information of common structure shared across different subpopulations, and thus can be inefficient in some real applications. For instance, in the glioblastoma multiforme (GBM) cancer dataset from The Cancer Genome Atlas Research Network (6), (7) showed that GBM cancer could be classified into four subtypes. Based on this cluster structure, it

has been suggested that although the graphs across four subtypes differ in some edges, they share many common structures. In this case, the naive procedure can be suboptimal (8; 9). Such applications have motivated recent studies on joint estimation methods (8; 9; 10; 11; 12; 13; 14) that encourage common structure in estimating heterogeneous graphical models. However, all aforementioned approaches crucially rely on an assumption that the class label of each sample is known in advance.

For certain problems, prior knowledge of the class membership may be available. But this may not be the case for the massive data with complex and unknown population structures. For instance, in online advertising, an important task is to find the most suitable advertisement (ad) for a given user in a specific online context. This could increase the chance of users' favorable actions (e.g., click the ad, inquire about or purchase a product). In recent years, user clustering has gained increasing attention due to its superior performance of ad targeting. This is because users with similar attributes, such as gender, age, income, geographic information, and online behaviors, tend to behave similarly to the same ad (15). Moreover, it is very important to understand conditional dependence relationships among user attributes in order to improve ad targeting accuracy (16). Such conditional dependence relationships are expected to share commonality across different groups (user homogeneity) while maintaining some levels of uniqueness within each group (user heterogeneity) (17). In this online advertising application, previously mentioned joint estimation methods are no longer applicable as they need to know the user cluster structure in advance. Furthermore, with the data being continuously collected, the number of underlying user clusters grows with the sample size (18). This provides another reason for simultaneously conducting user clustering and joint graphical model estimation, which is much needed in the era of big data.

Our contributions in this work are two-fold. On the methodological side, we propose a general framework of **S**imultaneous **C**lustering **A**nd estimatio**N** of heterogeneous graphical models (SCAN). SCAN is a likelihood based method which treats the underlying class label as a latent variable. Based on a high-dimensional version of

Expectation Conditional Maximization (ECM) algorithm (1), we are able to conduct clustering and sparse graphical model learning at the same time. In each iteration of the ECM algorithm, the expectation step performs cluster analysis by estimating missing labels and the conditional maximization step conducts feature selection and joint estimation of heterogeneous graphical models via a penalization procedure. With an iteratively updating process, the estimation for both cluster structure and sparse precision matrices becomes more and more refined. Our algorithm is computationally efficient by taking advantage of the fast sparse learning in the conditional maximization step. Moreover, it can be implemented in a user-friendly fashion, without the need of additional unsupervising learning knowledge.

As a promising application, we apply the SCAN method on the GBM cancer dataset to simultaneously cluster the GBM patients and construct the gene regulatory network of each subtype. Our method greatly outperforms the competitors in clustering accuracy and delivers new insights in understanding the GBM disease. Figure 1.1 reports four gene networks estimated from the SCAN method. The black lines are links shared in all four subtypes, and the color lines are uniquely presented in some subtypes. Our findings generally agree with the GBM disease literature (7). Besides common edges of all subtypes, we have discovered some unique gene connections that were not found through separate estimation (8; 9). This new finding suggests further investigation on their possible impact on the GBM disease. See Section 2.3.5 for more discussions.

On the theoretical side, we develop non-asymptotic statistical analysis for the output directly from the high dimensional ECM algorithm. This is nontrivial due to the non-convexity of the likelihood function. In this case, there is no guarantee that the sample-based estimator is close to the maximum likelihood estimator. Hence, we need to directly evaluate the estimation error in each iteration. Let $\mathbf{\Theta}$ represent vectorized cluster means $\boldsymbol{\mu}_k$ and precision matrices $\mathbf{\Omega}_k$, see (2.1.2) for a formal definition. Given

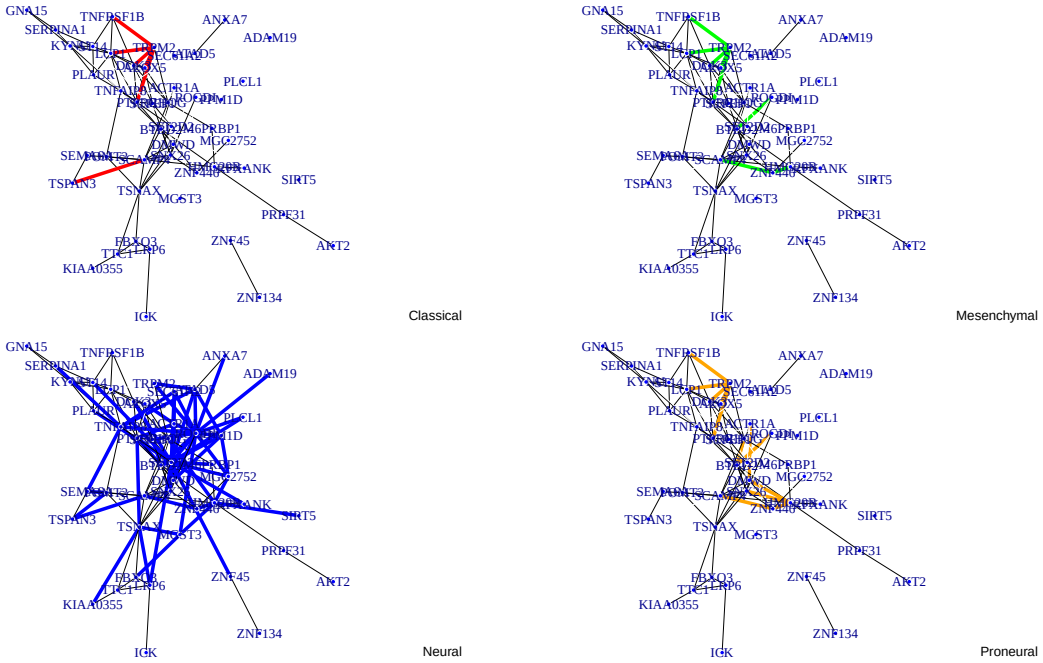

Fig. 1.1. Estimated gene networks corresponding to the Classical, Mesenchymal, Neural and Proneural clusters from our SCAN method applying to the Glioblastoma Cancer Data. In each network, the black lines are the links shared in all four groups. The color lines are the edges shared by some subtypes.

an appropriate initialization $\mathbf{\Theta}^{(0)}$, the finite sample error bound of the t-th step solution $\mathbf{\Theta}^{(t)}$ consists of two parts:

$$\left\|\mathbf{\Theta}^{(t)} - \mathbf{\Theta}^*\right\|_2 \leq \underbrace{C \cdot \varepsilon\left(n, p, K, \Psi(\mathcal{M})\right)}_{\text{Statistical Error(SE)}} + \underbrace{\kappa^t \left\|\mathbf{\Theta}^{(0)} - \mathbf{\Theta}^*\right\|_2}_{\text{Optimization Error(OE)}}, \tag{1.1.1}$$

with high probability. Here, K is the number of clusters, $\Psi(\mathcal{M})$ measures the sparsity of cluster means and precision matrices, and $\kappa \in (0, 1)$ is a contraction coefficient. The above theoretical analysis is applicable to any decomposable penalty used in the conditional maximization step.

The error bound (1.1.1) enables us to monitor the dynamics of estimation error in each iteration. Specifically, the optimization error decays geometrically with the iteration number t, while the statistical error remains the same when t grows. Therefore,

the maximal number of iterations T is implied, beyond which the optimization error is dominated by the statistical error such that consequently the whole error bound is in the same order as the statistical error. In particular,

$$\sum_{k=1}^{K}\left(\left\|\boldsymbol{\mu}_k^{(T)}-\boldsymbol{\mu}_k^*\right\|_2+\left\|\boldsymbol{\Omega}_k^{(T)}-\boldsymbol{\Omega}_k^*\right\|_F\right)=O_P\left(\underbrace{\sqrt{\frac{K^5 d\log p}{n}}}_{\text{Cluster means error}}+\underbrace{\sqrt{\frac{K^3(Ks+p)\log p}{n}}}_{\text{Precision matrices error}}\right),$$

where d and s are the sparsity for a single cluster mean and precision matrix. This result indicates that, after T steps, the SCAN estimator will fall within statistical precision of the true parameter $\{\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*\}$. It is worth mentioning that our theory allows the number of clusters K to diverge polynomially with the sample size, reflecting a typical big data scenario. When K is fixed, our statistical rate for the precision matrix estimation under the Frobenius norm, i.e., $O_P(\sqrt{(s+p)\log p/n})$, achieves the optimal rate established in Theorem 7 of (19), which is the best rate we could obtain even when the true cluster structure is given.

In the literature, a related line of research focuses on methodological developments of high-dimensional clustering. (20) and (21) introduced regularized model-based clustering and regularized K-means clustering, and (22) proposed a network-based clustering approach by imposing a graphical lasso to each individual precision matrix estimation. However, the regularized model-based clustering assumes an identical covariance matrix in each cluster, while the network-based clustering treats each graphical model estimation separately. As pointed out in (8) and (9), ignoring the network information of other clusters may lead to suboptimal graphical model estimation. During the submission of our paper, we became aware of an independent work by (23) who also considered the multiple precision matrices estimation via a Gaussian mixture model. Different from ours, (23) did not enforce the sparsity in the cluster means, which would inevitably lead to sub-optimal estimators in high-dimensional clustering (24; 25). Most importantly, no theoretical guarantee was provided in (22) and (23). On the other hand, our SCAN method is more general than these existing methods since we allow the sparsity in both cluster means and

precision matrices, and our theoretical analysis of the general SCAN framework sheds some lights on the behavior of these existing method, See Remark 2.1.1 for more discussions. In addition, in terms of the heterogeneous graphical model estimation, (26) proposed an interesting two-stage method which used hierarchical clustering to obtain cluster memberships and then estimated the multiple graphical models based on the attained cluster assignments. Despite its simplicity, it is unclear how the performance of clustering in the first stage could affect the performance of precision matrix estimation in the second stage. In comparison, our approach unifies clustering and parameter estimation into one optimization framework, which allows us to quantify both estimation errors in each iteration.

Another line of related work is the theoretical analysis of EM algorithm (24; 25; 27). Specifically, (27) studied the low-dimensional Gaussian mixture model, while (25) and (24) considered its high dimensional extensions. However, their methods are not applicable for the estimation of heterogeneous graphical models due to the assumed identity covariance matrix. In fact, our consideration of the general covariance matrix demands more challenging technical analysis since simultaneous estimation of cluster means and covariance matrices induces a bi-convex optimization beyond the non-convexity of the EM algorithm itself. This also explains why ECM is needed instead of EM. To address these technical issues, key ingredients of our theoretical analysis are to bound the dual norm of the gradient of an auxiliary Q-function and employ nice properties of bi-convex optimization (28) in the regularized M-estimation framework (29). See Section 2.2 for more details.

1.2 Sparse and Low-rank Tensor Estimation via Cubic Sketchings

The rapid advance in modern scientific technology gives rise to a wide range of high-dimensional tensor data (30; 31). Accurate estimation and fast communication/processing of tensor-valued parameters are crucially important in practice. For example, a tensor-valued predictor, which characterizes the association between brain

diseases and scientific measurements, such as magnetic resonance imaging, becomes the point of interest (32; 33; 34). Another example is tensor-valued image acquisition algorithms that can considerably reduce the number of required samples by exploiting the compressibility property of signals (35; 36).

In particular, the following tensor estimation model is widely considered in recent literatures,

$$y_i = \langle \mathscr{T}^*, \mathscr{X}_i \rangle + \epsilon_i, \quad i = 1, \ldots, n. \tag{1.2.1}$$

Here, $\mathscr{X}_i$ and ϵ_i are the measurement tensor and the noise, respectively. The goal is to estimate the unknown tensor $\mathscr{T}^*$ from measurements $\{y_i, \mathscr{X}_i\}_{i=1}^n$. A number of specific settings with varying forms of $\mathscr{X}_i$ have been studied, e.g., tensor completion (37; 38; 39; 40), tensor regression (32; 33; 34; 41; 42; 43), multi-task learning (44), etc.

In this work, we focus on the case that the measurement tensor can be written in a cubic form. For example, $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$ or $\mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i$, depending on $\mathscr{T}^*$ is symmetric or not. The cubic sketching form of $\mathscr{X}_i$ is motivated by a number of applications.

- *Interaction effect estimation:* High-dimensional high-order interaction models have been considered under a variety settings (45; 46; 47; 48). By writing $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$, we find that the interaction model has an interesting tensor representation (see left panel of Figure 1.2) which allows us to estimate high-order interaction terms using tensor techniques. This is in contrast with the existing literature that mostly focused on pair-wise interactions due to the model complexity and computational difficulties. More detailed discussions will be provided in Section 3.4.

- *High-order imaging/video compression:* High-order imaging/video compression is an important task in modern digital imaging with various applications (see right panel of Figure 1.2), such as hyper-spectral imaging analysis (49) and facial imaging recognition (50). In contrast to Gaussian ensembles for compression that each entry of $\mathscr{X}_i$ is i.i.d. randomly generated (32; 41; 42), the non-symmetric

cubic sketchings, i.e., $\mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i$, reduces the memory storage from $O(np^3)$ to $O(np)$, where n is sample size and p is the maximal dimension of tensor modes, but still preserve the optimal statistical rate. More detailed discussions will be provided in Section 3.5.

In practice, the total number of measurements n is considerably smaller than the number of parameters in unknown tensor $\mathscr{T}^*$, due to all kinds of restrictions such as time and storage. Fortunately, a variety of high-dimensional tensor data possess intrinsic structures, such as low-rankness (31) and sparsity (51), which highly reduce the effective dimension of the parameter and make the accurate estimation possible. Please refer to (3.2.2) and (3.5.2) for low-rank and sparse assumptions.

Fig. 1.2. Illustration for interaction reformulation and tensor image/video compression.

In this work, we propose a computationally efficient non-convex optimization approach for sparse and low-rank tensor estimation via cubic-sketchings. Our procedure is two-stage:

(i) obtain an initial estimate via the method of tensor moment (motivated by high-order Stein's identity), and then apply sparse tensor decomposition to the initial estimate to output a provably warm start;

(ii) use a thresholded gradient descent to iteratively refine the warm start along each tensor mode until convergence.

In theory, we carefully characterize the optimization and statistical errors at each iteration step. The output estimate is shown to converge in a geometric rate to an estimation with minimax optimal rate in statistical error (in terms of tensor Frobenius norm). In particular, after a logarithm factor of iterations, whenever $n \gtrsim K^2(s\log(ep/s))^{\frac{3}{2}}$, the proposed estimator $\widehat{\mathscr{T}}$ achieves

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \leq C\sigma^2 \frac{Ks\log(p/s)}{n}, \tag{1.2.2}$$

with high probability, where s, K, p, and σ^2 are the sparsity, rank, dimension, and noise level, respectively. We further establish the matching minimax lower bound to show that (1.2.2) is indeed optimal over a large class of sparse low-rank tensors. Our optimality result can be further extended to the non-sparse domain (such as tensor regression (42; 52)) – to the best of our knowledge, this is the first optimality result in both sparse and non-sparse low-rank tensor regressions.

The above theoretical analyses are non-trivial due to the non-convexity of the empirical risk function, and the need to develop some new high-order sub-Gaussian concentration inequalities. Specifically, the empirical risk function in consideration satisfies neither restricted strong convexity (RSC) condition nor sparse eigenvalue (SE) condition in general. Thus, many previous results, such as the one based on local optima analysis (42; 53; 54), are not directly applicable. Moreover, the structure of cubic-sketching tensor leads to high-order products of sub-Gaussian random variables. Thus, the matrix analysis based on Hoeffding-type or Bernstein-type concentration inequality (55; 56) will lead to sub-optimal statistical rate and sample complexity. This motivates us to develop new high-order concentration inequalities and sparse tensor-spectral-type bound, i.e., Lemmas 1 and 2 in Section 3.3.3. These new technical results are obtained based on the careful partial truncation of high-order products of sub-Gaussian random variables and the argument of bounded ψ_α-norm (57), and may be of independent interest.

A related line of research is low-rank matrix estimation in the literature, e.g., the spectral method and nuclear norm minimization (58; 59; 60). However, our cubic sketching model is by-no-means a simple extension from matrix estimation problems.

In general, many related concepts or methods for matrix data, such as singular value decomposition, are problematic to apply in the tensor framework (61; 62). It is also found that simple unfolding or matricizing of tensors may lead to suboptimal results due to the loss of structural information (63). Technically, the tensor nuclear norm is NP-hard to even approximate (38; 64), and thus the method to handle tensor low-rankness is particularly different from the matrix.

Notation Throughout the paper, vector, matrix, and tensor are denoted by boldface lower-case letters (e.g., $\boldsymbol{x}, \boldsymbol{y}$), boldface upper-case letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}$), and script letters (e.g., $\mathcal{X}, \mathcal{Y}$), respectively. For any set A, let $|A|$ be the cardinality. The $\mathrm{diag}(\boldsymbol{x})$ is a diagonal matrix generated by $\boldsymbol{x}$. For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{x} \circ \boldsymbol{y}$ is the outer product. Define $\|\boldsymbol{x}\|_q := (|x_1|^q + \cdots + |x_p|^q)^{1/q}$. We also define the l_0 quasi-norm by $\|\boldsymbol{x}\|_0 = \#\{j : x_j \neq 0\}$ and l_∞ norm by $\max_{1\leq j\leq p} |x_j|$. Let $\boldsymbol{e}_j$ be the canonical vectors, whose j-th entry equals to 1 and all other entries equal to zero. We use $[K]$ to denote the set $\{1, 2, \ldots, K\}$. For a vector $\boldsymbol{\mu} \in \mathbb{R}^p$, $\|\boldsymbol{\mu}\|_2$ is its Euclidean norm. For a matrix $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2}$, we denote $\|\boldsymbol{X}\|_F$ and $\|\boldsymbol{X}\|_2$ as its Frobenius norm and spectral norm, respectively, and define its matrix max norm as $\|\boldsymbol{X}\|_{\max} = \max_{i,j} |X_{ij}|$ and its max induced norm as $\|\boldsymbol{X}\|_\infty = \max_{i=1,\ldots,p_1} \sum_{j=1}^{p_2} |X_{ij}|$, which is simply the maximum absolute row sum of the matrix. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, let $\sigma_{\min}(\boldsymbol{A})$ and $\sigma_{\max}(\boldsymbol{A})$ be its smallest and largest eigenvalue respectively and $|\boldsymbol{A}|$ be its determinant. For a sub-Gaussian random variable Z, we use $\|Z\|_{\psi_2}$ and $\|Z\|_{\psi_1}$ to denote its Orlicz norm. Specifically, $\|Z\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|Z|^p)^{1/p}$ and $\|Z\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|Z|^p)^{1/p}$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, $a_n \lesssim b_n$ refers to the case that $a_n \leq C b_n$ for some uniform constant C. We write $\mathbb{1}(\cdot)$ as an indicator function. Throughout this thesis, we use $C, C_1, C_2, \ldots D, D_1, D_2, \ldots$ to denote generic absolute constants, whose values may vary at different places.

2. SIMULTANEOUS CLUSTERING AND ESTIMATION OF HETEROGENEOUS GRAPHICAL MODELS

2.1 Methodology

In this section, we introduce the SCAN method that simultaneously conducts high-dimensional clustering and estimation of heterogeneous graphical models.

2.1.1 Heterogeneous Graphical Models

We start our discussions from heterogeneous graphical models with known labels. Assume we are given K groups of data sets $\mathcal{A}_1, \ldots, \mathcal{A}_K$ and the samples in the k-th group are generated i.i.d. from the following Gaussian distribution:

$$f_k(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \right\}, k = 1, \ldots, K. \tag{2.1.1}$$

Let $\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k^{-1}$ be the k-th precision matrix with the ij-th entry ω_{kij}. For the k-th pair of parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$, i.e.,

$$\boldsymbol{\mu}_k = \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{pmatrix}, \boldsymbol{\Omega}_k = \begin{pmatrix} \omega_{k11} & \cdots & \omega_{k1p} \\ \vdots & \ddots & \vdots \\ \omega_{kp1} & \cdots & \omega_{kpp} \end{pmatrix},$$

we write $\boldsymbol{\Theta}_k := \vec{(}\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) = (\mu_{k1}, \ldots, \mu_{kp}, \omega_{k11}, \ldots, \omega_{kp1}, \ldots, \omega_{k1p}, \ldots, \omega_{kpp}) \in \mathbb{R}^{p^2+p}$ as its vectorized representation, and write the parameter of interest $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = \left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K\right)^\top \in \mathbb{R}^{K(p^2+p)}. \tag{2.1.2}$$

Note that the degrees of freedom of $\boldsymbol{\Theta}$ are $K(0.5p^2 + 1.5p)$, including K sets of p means, p variances, as well as $p(p-1)/2$ covariances.

In some cases, there may also exist some common structure across K precision matrices. (8) formulated the joint estimation of heterogeneous graphical models as

$$\underset{\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K\succ 0}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{\boldsymbol{x}\in\mathcal{A}_k} \log f_k(\boldsymbol{x};\boldsymbol{\Theta}_k) - \mathcal{P}(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K), \tag{2.1.3}$$

where $\mathcal{P}(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K)$ is an entry-wise penalty which encourages both sparsity of each individual precision matrix and similarity among all precision matrices.

In practice, the cluster label is not always available. A probabilistic model is thus needed to accommodate the latent structure in the data. Assume the observation $\boldsymbol{x}_i; i = 1,\ldots,n$, from unlabeled heterogeneous population has the underlying density

$$f(\boldsymbol{x},\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\Theta}_k), \tag{2.1.4}$$

where π_k is the probability that an observation $\boldsymbol{x}_i$ belongs to the k-th subpopulation. Here, for simplicity we assume the number of cluster K is identifiable. In order to ensure the identifiability of fixed-dimensional Gaussian graphical models, some sufficient conditions such as the strong identifiability condition was imposed on the density functions. However these conditions are hard to verify in practice. In fact, the identifiability issue for high dimensional mixture model is still an open problem (65) and is beyond the scope of this paper.

Consider the penalized log-likelihood function for the *observed data*

$$\log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}) := \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k\Big(\boldsymbol{x}_i;\boldsymbol{\mu}_k,(\boldsymbol{\Omega}_k)^{-1}\Big)\right) - \mathcal{R}(\boldsymbol{\Theta}).$$

Our **S**imultaneous **C**lustering **A**nd estimatio**N** (SCAN) method aims to solve

$$\max_{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Omega}_k} \log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}). \tag{2.1.5}$$

For an illustration, we take

$$\mathcal{R}(\boldsymbol{\Theta}) = \lambda_1 \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_{kj}|}_{\mathcal{P}_1(\boldsymbol{\Theta})} + \lambda_2 \underbrace{\sum_{k=1}^{K}\sum_{i\neq j}|\omega_{kij}|}_{\mathcal{P}_2(\boldsymbol{\Theta})} + \lambda_3 \underbrace{\sum_{i\neq j}(\sum_{k=1}^{K}\omega_{kij}^2)^{1/2}}_{\mathcal{P}_3(\boldsymbol{\Theta})}, \tag{2.1.6}$$

where $\mathcal{P}_1(\boldsymbol{\Theta})$ and $\mathcal{P}_2(\boldsymbol{\Theta})$ impose sparsity of the estimated cluster mean and precision matrix, and $\mathcal{P}_3(\boldsymbol{\Theta})$ encourages similarity among all estimated precision matrices. The above three tuning parameters can be tuned efficiently via adaptive BIC. More details can be found in Section 2.3.1.

Remark 2.1.1. It is worth mentioning that our SCAN method is applicable to penalty functions other than (2.1.6). For instance, the cluster mean penalty can be replaced by the group lasso penalty in (21) or the ℓ_0-norm penalty in (66). The group graphical lasso penalty for the precision matrix estimation can be substituted by the structural pursuit penalty in (67) or the weighted bridge penalty in (68). As shown in Section 2.1.2, only a slight modification of our algorithm is needed to accommodate other penalty functions. We also note that SCAN reduces to the regularized model-based clustering (20) when $\lambda_2 = \lambda_3 = 0$, reduces to the method by (22) when $\lambda_3 = 0$, and reduces to the method by (23) when $\lambda_1 = 0$. Consequently, the technical tools developed for the SCAN estimator in Section 2.2 are also applicable to these special cases.

2.1.2 ECM Algorithm

In this subsection, we introduce an efficient ECM algorithm to solve the general non-convex optimization problem in (2.1.5). The ECM replaces each M-step with an conditional maximization (CM) step in which each parameter $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k$ is maximized separately, by fixing other parameters.

Denote the latent cluster assignment matrix as $\boldsymbol{L}$, where $L_{ik} = \mathbb{1}(\boldsymbol{x}_i \in \mathcal{A}_k)$; $i = 1, \ldots, n$, $k = 1, \ldots, K$. If the cluster label L_{ik} is available, the penalized log-likelihood function for the *complete data* can be formulated as

$$\log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{L}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} L_{ik} \Big[\log \pi_k + \log f_k(\boldsymbol{x}_i; \boldsymbol{\Theta}_k) \Big] - \mathcal{R}(\boldsymbol{\Theta}).$$

In the expectation step, the conditional expectation of the penalized log-likelihood function is computed as

$$\mathbb{E}_{\boldsymbol{L}|\boldsymbol{X},\boldsymbol{\Theta}^{(t-1)}}\Big[\log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{L})\Big] = Q_n(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}) - \mathcal{R}(\boldsymbol{\Theta}), \tag{2.1.7}$$

where $\mathcal{R}(\boldsymbol{\Theta})$ is the penalty in (2.1.6) and

$$Q_n(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big[\log \pi_k + \log f_k(\boldsymbol{x}_i;\boldsymbol{\Theta}_k)\Big], \tag{2.1.8}$$

with the class label being computed based on the parameter $\boldsymbol{\Theta}^{(t-1)}$ and $\pi_k^{(t-1)}$ obtained at the previous iteration, that is,

$$L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i) = \frac{\pi_k^{(t-1)} f_k(\boldsymbol{x}_i;\boldsymbol{\Theta}_k^{(t-1)})}{\sum_{k=1}^{K}\pi_k^{(t-1)} f_k(\boldsymbol{x}_i;\boldsymbol{\Theta}_k^{(t-1)})}. \tag{2.1.9}$$

In the conditional maximization step, maximizing (2.1.7) with respect to π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Omega}_k$ yields the update of parameters. In particular, the update of π_k is given as

$$\pi_k^{(t)} = \sum_{i=1}^{n}\frac{L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)}{n}, \tag{2.1.10}$$

and the update of $\boldsymbol{\mu}_k$ is given in the following Lemma.

Lemma 2.1.2. Let $\boldsymbol{\mu}_k^{(t)} := \arg\max_{\boldsymbol{\mu}_k} Q_n(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}) - \mathcal{R}(\boldsymbol{\Theta})$ and denote $n_k := \sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)$. We have, for $j = 1,\ldots,p$,

$$\mu_{kj}^{(t)} = \begin{cases} g_{1,j}(\boldsymbol{x};\boldsymbol{\Theta}_k^{(t-1)}) - \frac{n\lambda_1}{n_k\omega_{kjj}^{(t-1)}}\operatorname{sign}(\mu_{kj}^{(t-1)}) & \text{if } \left|\sum_{i=1}^{n} g_{2,j}(\boldsymbol{x}_i;\boldsymbol{\Theta}_k^{(t-1)})\right| > \lambda_1; \\ 0 & \text{otherwise,} \end{cases}$$

where

$$g_{1,j}(\boldsymbol{x};\boldsymbol{\Theta}_k^{(t-1)}) = \frac{\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\big(\sum_{l=1}^{p} x_{il}\omega_{klj}^{(t-1)}\big)}{\omega_{kjj}^{(t-1)} n_k} - \frac{\sum_{l=1}^{p}\mu_{kl}^{(t-1)}\omega_{klj}^{(t-1)}}{\omega_{kjj}^{(t-1)}} + \mu_{kj}^{(t-1)},$$

$$g_{2,j}(\boldsymbol{x}_i;\boldsymbol{\Theta}_k^{(t-1)}) = L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big(\sum_{l=1,l\neq j}^{p}(x_{il}-\mu_{kl}^{(t-1)})\omega_{klj}^{(t-1)} + x_{ij}\omega_{kjj}^{(t-1)}\Big).$$

Note that if the lasso penalty is replaced with other penalty functions, then the update formula of $\boldsymbol{\mu}_k^{(t)}$ in Lemma 2.1.2 can be modified accordingly. Given the pseudo

sample covariance matrix $\widetilde{S}_k$, we are able to develop an update formula for $\mathbf{\Omega}_k$ by establishing its connection with joint estimation of heterogeneous graphical models (2.1.3).

Lemma 2.1.3. The solution of maximizing (2.1.7) with respect to $(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K)$ is equivalent to

$$(\mathbf{\Omega}_1^{(t)}, \ldots, \mathbf{\Omega}_K^{(t)}) := \arg \max_{\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K \succ 0} \sum_{k=1}^{K} n_k \Big[\log\det(\mathbf{\Omega}_k) - \mathrm{trace}(\widetilde{S}_k \mathbf{\Omega}_k) \Big] - \mathcal{R}(\mathbf{\Theta}), \quad (2.1.11)$$

where $\widetilde{S}_k$ is a pseudo sample covariance matrix defined as

$$\widetilde{S}_k := \frac{\sum_{i=1}^{n} L_{\mathbf{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t-1)})^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t-1)})}{\sum_{i=1}^{n} L_{\mathbf{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)}.$$

The solution for (2.1.11) can be solved efficiently via the ADMM algorithm by slightly modifying the joint graphical lasso algorithm in (8). Since (8) do not impose the symmetry condition for precision matrix update, $\{\mathbf{\Omega}_k^{(T)}\}_{k=1}^K$ in general is not necessarily symmetric. Following the symmetrization strategy in (69) and (13), we symmetrize $\mathbf{\Omega}_k^{(t)}$ by

$$\omega_{kij}^{(t)} = \omega_{kij}^{(t)} I(|\omega_{kij}^{(t)} \leq \omega_{kij}^{(t)}|) + \omega_{kji}^{(t)} I(|\omega_{kij}^{(t)} > \omega_{kij}^{(t)}|), \quad (2.1.12)$$

where $\omega_{kij}^{(t)}$ is the ij-th entry of $\mathbf{\Omega}_k^{(t)}$ and $I(\cdot)$ is the indicator function. This step will not affect the convergence rate of the final estimator, which is illustrated in (69) and (13). We summarize the high-dimensional ECM algorithm for solving the SCAN method in Table 2.1. Our algorithm is computationally efficient due to fast sparse learning routines shown in Lemmas 2.1.2 and 2.1.3.

In all of our experiments, we obtain $(\boldsymbol{\mu}_k^{(0)}, \mathbf{\Omega}_k^{(0)})$ by random initialization, which is computationally efficient and practically reliable. In the theoretical study, we require the initialization to be of a constant distance to the truth. See Remark 2.2.8 for more discussions. Moreover, in the implementation, ECM step in Step 2 is terminated when the updated parameters are close to their previous values:

$$\sum_{k=1}^{K} \left\{ \frac{\|\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^{(t-1)}\|_2}{\|\boldsymbol{\mu}_k^{(t)}\|_2} + \frac{\|\mathbf{\Omega}_k^{(t)} - \mathbf{\Omega}_k^{(t-1)}\|_F}{\|\mathbf{\Omega}_k^{(t)}\|_F} \right\} \leq 0.01.$$

Table 2.1.
The SCAN Algorithm

Input: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, number of clusters K, tuning parameters $\lambda_1, \lambda_2, \lambda_3$.
Output: Cluster label $\boldsymbol{L}$, cluster mean $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Omega}_k$.
Step 1: Initialize cluster mean $\boldsymbol{\mu}_k^{(0)}$, positive definite precision matrix $\boldsymbol{\Omega}_k^{(0)}$, and set $\pi_k^{(0)} = 1/K$, for each $k \in [K]$.
Step 2: Until some termination conditions are met, for iteration $t = 1, 2, \ldots$
(a) E-step. Find the cluster assignment $L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)$ as in (2.1.9).
(b) CM-step. Given $L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)$, update $\pi_k^{(t)}$, $\boldsymbol{\mu}_k^{(t)}$, and $\boldsymbol{\Omega}_k^{(t)}$ in (2.1.10), Lemma 2.1.2, Lemma 2.1.3, respectively. Symmetrize $\boldsymbol{\Omega}_k^{(t)}$ by (2.1.12).

Remark 2.1.4. In the existing high-dimensional EM algorithms where the covariance matrix is assumed to be an identity matrix (24; 25), sample-splitting procedures have been routinely used in the M-step in order to facilitate the theoretical analysis. Although it simplifies theoretical developments, such a sample-splitting procedure does not take advantage of full samples in the M-step and is hard to implement in practice. Our Algorithm 2.1 is able to avoid this sample-splitting step but still enjoys nice theoretical properties. See Corollary 2.2.12 for more discussions on its statistical guarantee.

2.2 Statistical Guarantee

In this section, we establish statistical guarantee for the SCAN estimator based on sample-based analysis of (2.1.8) and population-based analysis of (2.2.3). Here, we consider the high-dimensional setting where $p \gg n$ and K is allowed to diverge with n.

We start by introducing some useful notation. Denote the index set of diagonal components of K precision matrices by

$$\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}_k, \text{with } \mathcal{G}_k = \big(k(p+1), k(2p+2), \ldots, k(p^2+p)\big), \tag{2.2.1}$$

that is, $\boldsymbol{\Theta}_{\mathcal{G}} = (\omega_{111}, \ldots, \omega_{1pp}, \ldots, \omega_{K11}, \ldots, \omega_{Kpp}) \in \mathbb{R}^{Kp}$. Let $\mathcal{O}$ be the complete index set of $\boldsymbol{\Theta}$ and $\mathcal{G}^c = \mathcal{O} \setminus \mathcal{G}$ be the complement set of $\mathcal{G}$. Denote $\mathcal{U}_k := \{i : \mu^*_{ki} \neq 0\}$ where $\boldsymbol{\mu}^*_k$ is the true mean parameter, $\mathcal{V}_k := \{(i,j) : i \neq j, \omega^*_{kij} \neq 0\}$ where $\boldsymbol{\Omega}^*_k$ is the true precision matrix and $\mathcal{S}_1 = \bigcup_{k=1}^{K} \mathcal{U}_k$, $\mathcal{S}_2 = \bigcup_{k=1}^{K} \mathcal{V}_k$. Define $\Xi \subseteq \mathbb{R}^{K(p^2+p)}$ as some non-empty convex set of parameters. Denote the support space $\mathcal{M}$ as

$$\begin{aligned} \mathcal{M} \;:=\; \Big\{ \boldsymbol{V} \in \Xi \;\Big|\; & \mu_{ki} = 0 \text{ for all } i \notin \mathcal{S}_1, \\ & \omega_{kij} = 0 \text{ for all pairs } (i,j) \notin \mathcal{S}_2, k = 1 \ldots, K \Big\}, \end{aligned} \tag{2.2.2}$$

where $\boldsymbol{V}$ follows the same definition style used for $\boldsymbol{\Theta}$ in (2.1.2). Denote the sparsity parameters:

$$\begin{aligned} s &:= \#\{(i,j) : \omega^*_{kij} \neq 0, i, j = 1 \ldots p, i \neq j, k = 1, \ldots, K\}, \\ d &:= \#\{i : \mu^*_{ik} \neq 0, i = 1, \ldots, p, k = 1, \ldots, K\}. \end{aligned}$$

2.2.1 Population-Based Analysis

We define a corresponding population version of Q_n in (2.1.8) as

$$Q(\boldsymbol{\Theta}^{'} | \boldsymbol{\Theta}) := \mathbb{E}\left[\sum_{k=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) [\log \pi^{'}_k + \log f_k(\boldsymbol{X}; \boldsymbol{\Theta}^{'}_k)] \right]. \tag{2.2.3}$$

Without loss of generality, we assume the true prior probability $\pi^*_k = 1/K$ for each $k = 1, \ldots, K$. Recall that the update of weights in (2.1.10) is independent of the updates of other parameters. Consequently, according to (2.1.1), maximizing $Q(\boldsymbol{\Theta}^{'} | \boldsymbol{\Theta})$ over $(\boldsymbol{\mu}^{'}_k, \boldsymbol{\Omega}^{'}_k)$ is equivalent to maximizing

$$\sum_{k=1}^{K} \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \left\{ \frac{1}{2} \log\det(\boldsymbol{\Omega}^{'}_k) - \frac{1}{2} (\boldsymbol{X} - \boldsymbol{\mu}^{'}_k)^{\top} \boldsymbol{\Omega}^{'}_k (\boldsymbol{X} - \boldsymbol{\mu}^{'}_k) \right\} \right]. \tag{2.2.4}$$

Clearly, the update of $(\boldsymbol{\mu}_l^{'}, \boldsymbol{\Omega}_l^{'})$ is independent of the update of $(\boldsymbol{\mu}_t^{'}, \boldsymbol{\Omega}_t^{'})$ for any $t \neq l$. This enables us to characterize the update of each pair of parameters separately. For any $k = 1, \ldots, K$, define

$$M_{\boldsymbol{\mu}_k^{'}}(\boldsymbol{\Omega}_k^{'}) := \arg\max_{\boldsymbol{\mu}_k^{'}} Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta}) \text{ and } M_{\boldsymbol{\Omega}_k^{'}}(\boldsymbol{\mu}_k^{'}) := \arg\max_{\boldsymbol{\Omega}_k^{'}} Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta}).$$

We show in Lemma 2.2.1 that the population update of $\boldsymbol{\mu}_k^{'}$ is independent of $\boldsymbol{\Omega}_k^{'}$, while the population update of $\boldsymbol{\Omega}_k^{'}$ is a function of $\boldsymbol{\mu}_k^{'}$.

Lemma 2.2.1. For any $k = 1, \ldots, K$, we have

$$M_{\boldsymbol{\mu}_k^{'}}(\boldsymbol{\Omega}_k^{'}) = \left[\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\right]^{-1}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}], \tag{2.2.5}$$

$$M_{\boldsymbol{\Omega}_k^{'}}(\boldsymbol{\mu}_k^{'}) = \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\left[\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X}-\boldsymbol{\mu}_k^{'})(\boldsymbol{X}-\boldsymbol{\mu}_k^{'})^{\top}]\right]^{-1}. \tag{2.2.6}$$

The difficulty of simultaneous clustering and estimation can be characterized by the following *sufficiently separable condition*. Define $\mathcal{B}_\alpha(\boldsymbol{\Theta}^*) := \left\{\boldsymbol{\Theta} \in \Xi : \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_2 \leq \alpha\right\}$.

Condition 2.2.2 (*Sufficiently Separable Condition*)**.** Denote $W = \max_j W_j$, $W^{'} = \max_j W_j^{'}$, $W^{''} = \max_j W_j^{''}$ with $W_j, W_j^{'}, W_j^{''}$ defined in (2.6.4), (2.6.7) and (2.6.8), respectively. We assume K clusters are sufficiently separable such that given an appropriately small parameter $\gamma > 0$, it holds a.s.

$$L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \cdot L_{\boldsymbol{\Theta},j}(\boldsymbol{X}) \leq \frac{\gamma}{24(K-1)\sqrt{\max\{W, W^{'}, W^{''}\}}}, \tag{2.2.7}$$

for each pair $\{(j,k), j,k \in [K], j \neq k\}$ and any $\boldsymbol{\Theta} \in \mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$.

Condition 2.2.2 requires that K clusters are sufficiently separable in the sense that $\boldsymbol{X}$ belongs to the k-th cluster with probability either close to zero or close to one such that $L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \cdot L_{\boldsymbol{\Theta},j}(\boldsymbol{X})$ is close to zero. In the special case that $K = 2$ and $\boldsymbol{\Omega}_1^* = \boldsymbol{\Omega}_2^* = \mathbb{1}_p$, (27) requires $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2$ is sufficiently large. Our Condition 2.2.2 extends it to general K and general precision matrices. Note that the condition (2.2.7) is related with the number of clusters K. As K grows, the clustering problem gets harder and hence a stronger sufficiently separable condition is needed.

The next lemma guarantees that the curvature of $Q(\cdot|\boldsymbol{\Theta})$ is similar to that of $Q(\cdot|\boldsymbol{\Theta}^*)$ when $\boldsymbol{\Theta}$ is close to $\boldsymbol{\Theta}^*$, which is a key ingredient in our population-based analysis.

Lemma 2.2.3 (*Gradient Stability*)**.** Under Condition 2.2.2, the function $\{Q(\cdot|\boldsymbol{\Theta}), \boldsymbol{\Theta} \in \Xi\}$ satisfies,

$$\left\|\nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^*)\right\|_2 \leq \tau \cdot \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_2, \tag{2.2.8}$$

with parameter $\tau \leq \gamma/12$ for any $\boldsymbol{\Theta} \in \mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$. The gradient $\nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})$ is taken with respect to the first variable of $Q(\cdot|\cdot)$.

2.2.2 Sample-Based Analysis

In this section, we analyze the sample-base function Q_n, defined as the objective function in (2.1.8). The statistical error comes from the approximation by using sample-base function Q_n to population-base function Q. We need one regularity condition to ensure that Q_n is strongly concave in a specific Euclidean ball.

Condition 2.2.4. There exist some positive constants β_1, β_2 such that $0 < \beta_1 < \min_{k\in[K]} \sigma_{\min}(\boldsymbol{\Omega}_k^*) < \max_{k\in[K]} \sigma_{\max}(\boldsymbol{\Omega}_k^*) < \beta_2$.

Lemma 2.2.5 verifies the restricted strong concavity condition of Q_n. Note that (2.2.9) corresponds to the restricted eigenvalue condition in sparse linear regression (29).

Lemma 2.2.5 (*Restricted Strong Concavity*)**.** Suppose that Condition 2.2.4 holds. Then for any $\boldsymbol{\Theta} \in \mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$, with probability at least $1-\delta$, each $\boldsymbol{\Theta}' \in \mathbb{C} := \{\boldsymbol{\Theta}' \mid \|\boldsymbol{\Theta}' - \boldsymbol{\Theta}^*\|_2 \leq 2\alpha\}$ satisfies

$$Q_n(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) - Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \left\langle \nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}), \boldsymbol{\Theta}' - \boldsymbol{\Theta}^* \right\rangle \leq -\frac{\gamma}{2}\left\|\boldsymbol{\Theta}' - \boldsymbol{\Theta}^*\right\|_2^2, \tag{2.2.9}$$

with sufficiently large n, where $\gamma = c \cdot \min\{\beta_1, 0.5(\beta_2 + 2\alpha)^{-2}\}$ is the strong concavity parameter for some constant c.

Define $\mathcal{P}(\mathbf{\Theta}) = M_1\mathcal{P}_1(\mathbf{\Theta}) + M_2\mathcal{P}_2(\mathbf{\Theta}) + M_3\mathcal{P}_3(\mathbf{\Theta})$ for some positive constants M_1, M_2, M_3. Let $\mathcal{P}^*$ be the dual norm of $\mathcal{P}$, which is defined as $\mathcal{P}^*(\mathbf{\Theta}) = \sup_{\mathcal{P}(\mathbf{\Theta}')\leq 1}\langle \mathbf{\Theta}', \mathbf{\Theta}\rangle$. For simplicity, write $\|\cdot\|_{\mathcal{P}^*} = \mathcal{P}^*(\cdot)$.

Condition 2.2.6. For any fixed $\mathbf{\Theta} \in \mathcal{B}_\alpha(\mathbf{\Theta}^*)$, with probability at least $1-\delta_1$,

$$\Big\|\nabla Q_n(\mathbf{\Theta}^*|\mathbf{\Theta}) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta})\Big\|_{\mathcal{P}^*} \leq \varepsilon_1, \tag{2.2.10}$$

and with probability at least $1-\delta_2$, we have

$$\Big\|\Big[\nabla Q_n(\mathbf{\Theta}^*|\mathbf{\Theta}) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta})\Big]_{\mathcal{G}}\Big\|_2 \leq \varepsilon_2, \tag{2.2.11}$$

where $\mathcal{G}$ is the diagonal index set defined in (2.2.1). Here ε_1 and ε_2 are functions of $n, p, K, \delta_1, \delta_2$.

Intuitively, ε_1 and ε_2 quantify the difference between the population-based and sample-based conditional maximization step. Note that $\mathcal{P}$ does not penalize diagonal elements of each precision matrix, thus

$$\Big\|\nabla Q_n(\mathbf{\Theta}^*|\mathbf{\Theta}) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta})\Big\|_{\mathcal{P}^*} = \Big\|\Big[\nabla Q_n(\mathbf{\Theta}^*|\mathbf{\Theta}) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta})\Big]_{\mathcal{G}^c}\Big\|_{\mathcal{P}^*}.$$

Our analysis makes use of the property of dual norm to bridge the SCAN penalty term and the targeted error term in L_2 norm. Note that our SCAN penalty does not penalize diagonal terms of precision matrices, and hence it can be treated as a norm only if it is applied to the parameter $\mathbf{\Theta}$ without diagonal terms of precision matrices. Otherwise, it is a semi-norm. For this purpose, we separate all the diagonal terms from $\mathbf{\Theta}$. Therefore, our statistical error is split by two parts: one from the sparse estimate of cluster means and non-diagonal terms in precision matrices, and another from the estimate of diagonal terms of precision matrices. In Lemma 2.6.1, ε_1 and ε_2 will be specifically calculated for our proposed SCAN penalty. In the high dimensional ECM algorithm, there is no explicit form for the CM-step update due to the existence of the penalty term. This is a crucial difference from the low-dimensional EM algorithm in (27). Fortunately, the decomposability of SCAN penalty enables us to quantify statistical errors by evaluating the gradient of Q-function.

2.2.3 Statistical Error versus Optimization Error

In this section, we provide the final theoretical guarantee for the high-dimensional ECM algorithm by combining the population and sample-based analysis.

Definition 1 *Support Space Compatibility Constant*. For the support subspace $\mathcal{M} \subseteq \mathbb{R}^{K(p^2+p)}$ defined in (2.2.2), we define

$$\nu(\mathcal{M}) = \sup_{\mathbf{\Theta} \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{P}(\mathbf{\Theta})}{\|\mathbf{\Theta}\|_2}. \tag{2.2.12}$$

Remark 2.2.7. The support space compatibility constant $\nu(M)$ is a variant of subspace compatibility constant originally proposed by (29) and (70). Actually, $\nu(M)$ can be interpreted as a notion of intrinsic dimensionality of M. In order to bound the statistical error, we need some measures for the complexity of parameter $\mathbf{\Theta}$ reflected by the penalty term. One possible way is to specify a model subspace $\mathcal{M}$ and require $\mathbf{\Theta}$ lie in the space. By choosing the support space $\mathcal{M}$ of parameter of interest $\mathbf{\Theta}$, the support space compatibility constant $\nu(\mathcal{M})$ can measure the complexity of $\mathbf{\Theta}$ relative to the penalty term $\mathcal{P}$ and square norm. The larger $\nu(\mathcal{M})$ is, the more samples are needed to guarantee statistical consistency. For examples, if the penalty $\mathcal{P}$ is L_1 penalty with s-sparse coordinate support space $\mathcal{M}'$, then we have $\nu(\mathcal{M}') = \sqrt{s}$. In the context of group lasso penalty, we have $\nu(\mathcal{M}') = \sqrt{|S|}$, where S is the index set of active groups. For our SCAN penalty, $\nu(\mathcal{M})$ is specifically calculated by $M_1\sqrt{Kd}+(M_2\sqrt{K}+M_3)\sqrt{s}$, where d, s are the common sparsity parameters for single cluster means and precision matrices accordingly and M_1, M_2, M_3 are some absolute constants.

We first provide a general theory that applies to any decomposable penalty, such as the group lasso penalty in (21) and fused graphical lasso penalty in (8). The theoretical result of our SCAN penalty will be discussed in Corollary 2.2.12.

Theorem 1 . Suppose Conditions 2.2.2, 2.2.4, 2.2.6 hold and $\mathbf{\Theta}^*$ lies in the interior of Ξ. Let $\kappa = 6\tau/\gamma$, where τ, γ are calculated in Lemma 2.2.3 and Lemma 2.2.5. Consider our SCAN algorithm in Table 2.1 with initialization $\mathbf{\Theta}^{(0)}$ falling into a ball

$\mathcal{B}_\alpha(\mathbf{\Theta}^*)$ for some constant radius $\alpha > 0$ and assume the tuning parameters satisfy $\lambda_1 = M_1\lambda_n^{(t)}$, $\lambda_2 = M_2\lambda_n^{(t)}$, $\lambda_3 = M_3\lambda_n^{(t)}$, and

$$\lambda_n^{(t)} = \varepsilon + \kappa\frac{\gamma}{\nu(\mathcal{M})}\left\|\mathbf{\Theta}^{(t-1)} - \mathbf{\Theta}^*\right\|_2. \tag{2.2.13}$$

If the sample size n is large enough such that $\varepsilon \leq (1-\kappa)\gamma\alpha/(6\nu(\mathcal{M}))$, then $\mathbf{\Theta}^{(t)}$ satisfies, with probability at least $1 - t\delta'$,

$$\left\|\mathbf{\Theta}^{(t)} - \mathbf{\Theta}^*\right\|_2 \leq \underbrace{\frac{6\nu(\mathcal{M})}{(1-\kappa)\gamma}\varepsilon}_{\text{Statistical Error(SE)}} + \underbrace{\kappa^t\left\|\mathbf{\Theta}^{(0)} - \mathbf{\Theta}^*\right\|_2}_{\text{Optimiation Error(OE)}}, \tag{2.2.14}$$

where $\delta' = \delta + \delta_1 + \delta_2$ with δ, δ_1, δ_2 defined in Lemma 2.2.5 and Condition 2.2.6 and $\varepsilon = \varepsilon_1 + \varepsilon_2/\nu(\mathcal{M})$.

The above theoretical result suggests that the estimation error in each iteration consists *statistical error* and *optimization error*. From the definition of τ in Lemma 2.2.3, κ is less than 0.5 so that it is a contractive parameter. With a relatively good initialization, even though ECM algorithm may be trapped into a local optima after enough iterations, it can be guaranteed to be within a small neighborhood of the truth, in the sense of statistical accuracy. In addition, with a proper choice of δ', the final probability $1 - t\delta'$ will converge to 1; see Corollary 2.2.12 for details.

Remark 2.2.8. To our limited knowledge, there is no existing literature to guarantee the global convergence of ECM algorithm in a general case. Compromisingly, we have to require some constraints on the initial value. In our framework, the only requirement for the initial value is to fall into a ball with constant radius to the truth. Such a condition has also been imposed in EM algorithms (24; 25; 27) and can be fulfilled by some spectral-based initializations (71).

Remark 2.2.9. In Theorem 1, we introduce an iterative turning procedure (2.2.13) which appeared in high dimensional regularized M-estimation (29), and was also applied in (24) to facilitate their theoretical analysis.

The error bound in (2.2.14) measures the estimation error in each iteration. Here, optimization error decays geometrically with the iteration number t, while

the statistical error remains the same when t grows. Therefore, this enables us to provide a meaningful choice of the maximal number of iterations T beyond which the optimization error is dominated by the statistical error such that the whole error bound is in the same order of the statistical error.

In the following corollary, taking the SCAN penalty as an example, we provide a closed form of the maximal number of iterations T and also an explicit form of the estimation error.

Condition 2.2.10. The largest element of cluster means and precision matrices are both bounded, that is, for some positive constants c_1 and c_2,

$$\|\boldsymbol{\mu}^*\|_\infty := \max_{k\in[K]} \|\boldsymbol{\mu}_k^*\|_\infty < c_1 \text{ and } \|\boldsymbol{\Omega}^*\|_{\max} := \max_{k\in[K]} \|\boldsymbol{\Omega}_k^*\|_{\max} < c_2.$$

Condition 2.2.11. Suppose that the number of clusters K satisfies $K^2 = o(p(\log n)^{-1})$.

Corollary 2.2.12. Suppose Conditions 2.2.2, 2.2.4, 2.2.10 and 2.2.11 hold. If sample size n is sufficiently large such that

$$n \geq \left(\frac{6(CK\|\boldsymbol{\Omega}^*\|_\infty + C'K^{1.5})(\sqrt{Kd} + \sqrt{Ks} + \sqrt{K}) + C''K^{1.5}\sqrt{p}}{(1-\kappa)\gamma\alpha} \right)^2 \log p,$$

and the iteration step t is large enough such that

$$t \geq T = \log_{1/\kappa} \frac{\left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2}{\varphi(n,p,K)},$$

where $\varphi(n,p,K) = 6\widetilde{C}((1-\kappa)\gamma)^{-1}\|\boldsymbol{\Omega}^*\|_\infty(\sqrt{Kd} + \sqrt{Ks+p})\sqrt{K^3 \log p/n}$ for some positive constant $\widetilde{C}$, the optimization error in (2.2.14) is dominated by the statistical error, and

$$\begin{aligned}
&\sum_{k=1}^K \left(\left\|\boldsymbol{\mu}_k^{(T)} - \boldsymbol{\mu}_k^*\right\|_2 + \left\|\boldsymbol{\Omega}_k^{(T)} - \boldsymbol{\Omega}_k^*\right\|_F \right) \\
&\leq \frac{12\widetilde{C}}{(1-\kappa)\gamma} \left(\underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{K^5 d \log p}{n}}}_{\text{Cluster means error}} + \underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{K^3(Ks+p)\log p}{n}}}_{\text{Precision matrices error}} \right),
\end{aligned}$$

with probability converging to 1.

Remark 2.2.13. If K is fixed, the above upper bound reduces to

$$\begin{aligned}&\sum_{k=1}^{K}\left(\left\|\boldsymbol{\mu}_k^{(T)}-\boldsymbol{\mu}_k^*\right\|_2+\left\|\boldsymbol{\Omega}_k^{(T)}-\boldsymbol{\Omega}_k^*\right\|_F\right)\\&\lesssim\left(\underbrace{\|\boldsymbol{\Omega}^*\|_\infty\sqrt{\frac{d\log p}{n}}}_{\text{Cluster means error}}+\underbrace{\|\boldsymbol{\Omega}^*\|_\infty\sqrt{\frac{(s+p)\log p}{n}}}_{\text{Precision matrices error}}\right).\end{aligned}\tag{2.2.15}$$

Consider the class of precision matrix $\mathcal{Q}:=\{\boldsymbol{\Omega}:\boldsymbol{\Omega}\succ 0,\|\boldsymbol{\Omega}\|_\infty\le C_{\mathcal{Q}}\}$ as in (19). When $C_{\mathcal{Q}}$ does not depend on n,p, our rate $\sqrt{(s+p)\log p/n}$ in (2.2.15) is minimax optimal for estimating s-sparse precision matrix under Frobenius norm (see Theorem 7 in (19)). The same rate has also been obtained in (26) for multiple precision matrix estimation when the true cluster structure is assumed to be given in advance. Moreover, our cluster mean error rate $\sqrt{d\log p/n}$ is minimax optimal for estimating d-sparse cluster means; see (25). In short, Corollary 2.2.12 indicates that our procedure is able to achieve optimal statistical rates for both cluster means and multiple precision matrices even when the true cluster structure is unknown.

Remark 2.2.14. As a by-product, we establish the variable selection consistency of $\boldsymbol{\Omega}_k^{(T)}$, which ensures that our precision matrix estimator can asymptotically identify true connected links. Assume $\|\boldsymbol{\Omega}_k^*\|_\infty$ is bounded and the minimal signal in the true precision matrix satisfies $\omega_{\min}:=\min_{(i,j)\in\mathcal{V}_k,k=1,\ldots,K}w_{kij}^*>2r_n$, where $r_n=(\sqrt{K^5d}+\sqrt{K^3(Ks+p)})\sqrt{\log p/n}$. The latter condition is weaker than that assumed in (10), where they require a constant lower bound of $\omega_{\min}$. To ensure the model selection consistency, we threshold the precision matrix estimator $\boldsymbol{\Omega}_k^{(T)}$ such that $\widetilde{\omega}_{kij}=\omega_{kij}^{(T)}\mathbb{1}\{|\omega_{kij}^{(T)}|>r_n\}$ as in (72) and (9). See Theorem 2 in the online supplementary for some results on the selection consistency result.

2.3 Numerical Study

In this section, we discuss an efficient tuning parameter selection procedure and demonstrate the superior numerical performance of our method. We compare our algorithm with three clustering and graphical model estimation methods:

- Standard K-means clustering (73).
- Algorithm in (22) which applies graphical lasso for each precision matrix estimation.
- A two-stage approach which first uses K-means clustering to obtain the clusters and then applies joint graphical lasso (8) to estimate precision matrices.

For a fair comparison, we assume the number of clusters K is given in all methods.

2.3.1 Selection of Tuning Parameters

In our simultaneous clustering and graph estimation formulation, three tuning parameters $\Lambda := \{\lambda_1, \lambda_2, \lambda_3\}$ need to be appropriately determined so that both the clustering and network estimation performance can be optimized. In our framework, the tuning parameters are selected through the following adaptive BIC-type selection criterion. For a set of tuning parameters $\Lambda := \{\lambda_1, \lambda_2, \lambda_3\}$, the adaptive BIC criterion is defined as

$$\mathrm{BIC}(\Lambda) = -2 \log \widehat{L}(\Lambda) + \log(n) \mathrm{df}_\Lambda(\boldsymbol{\mu}) + 2\mathrm{df}_\Lambda(\boldsymbol{\Omega}), \tag{2.3.1}$$

where $\widehat{L}(\Lambda)$ is the sample likelihood function and $\{\mathrm{df}_\Lambda(\boldsymbol{\mu}), \mathrm{df}_\Lambda(\boldsymbol{\Omega})\}$ is the degrees of freedom of the model. Here, $\{\mathrm{df}_\Lambda(\boldsymbol{\mu}), \mathrm{df}_\Lambda(\boldsymbol{\Omega})\}$ can be approximated by the size of selected variables in the final estimator. Therefore, according to the Gaussian mixture model assumption, the adaptive BIC criterion in (2.3.1) can be computed as

$$-2 \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \widehat{\pi}_k f_k \Big(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_k, (\widehat{\boldsymbol{\Omega}}_k)^{-1} \Big) \right) + \sum_{k=1}^{K} \left\{ \log n \cdot s_{1k} + 2 s_{2k} \right\},$$

where $s_{1k} = \mathrm{Card}\{i : \widehat{\mu}_{ki} \neq 0\}$, $s_{2k} = \mathrm{Card}\{(i,j) : \widehat{\Omega}_{kij} \neq 0, 1 \leq i < j \leq p\}$ and $\widehat{\pi}_k, \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Omega}}_k$ are final updates from Algorithm 2.1. We choose a smaller weight for the degrees of freedom of precision matrices as suggested in (8). The mixing weight π is not counted into the degrees of freedom since it only contributes a constant factor.

In our experiment, we choose the optimal set of parameters minimizing the BIC value in (2.3.1). In the high-dimensional scenario where p is very large, calculation of

BIC over a grid search for all $\lambda_1, \lambda_2, \lambda_3$ may be computationally expensive. Following (8), we suggest a line search over λ_1, λ_2 and λ_3. In detail, we fix λ_2 and λ_3 at their median value of the given range and conduct a grid search over λ_1. Then with tuned λ_1 and median value of λ_3, we conduct a grid search over λ_2. The line search for λ_3 is the same. In our simulations, we choose the tuning range $10^{-2+2t/15}$ with $t = 0, 1, \ldots, 15$ for all $\lambda_1, \lambda_2, \lambda_3$.

2.3.2 Illustration

In this subsection, we demonstrate the importance of simultaneous clustering and estimation in improving both the clustering performance and the estimation accuracy of multiple precision matrices.

The simulated data consists of $n = 1000$ observations from 2 clusters, and among them 500 observations are from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and the rest 500 observations are from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}_1 = (0, 1)^\top$, $\boldsymbol{\mu}_2 = (0, -1)^\top$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The standard K-means algorithm treats the data space as isotropic (distances unchanged by translations and rotations) (74). This means that data points in each cluster are modeled as lying within a sphere around the cluster centroid. A sphere has the same radius in each dimension. However, the non-diagonal covariance matrix in the mixture model makes the cluster structure highly non-spherical. Thus, the K-means algorithm is expected to produce an unsatisfactory clustering result. This is illustrated in Figure 2.3.2 where K-means clustering clearly obtains wrong clusters. On the other hand, by incorporating the precision matrix estimation into clustering, our method is able to identify two correct clusters.

Figure 2.3.2 illustrates the estimation performance of precision matrices based on the clusters estimated from the K-means clustering and our method. Clearly, our SCAN method delivers an estimator with improved accuracy when compared to the

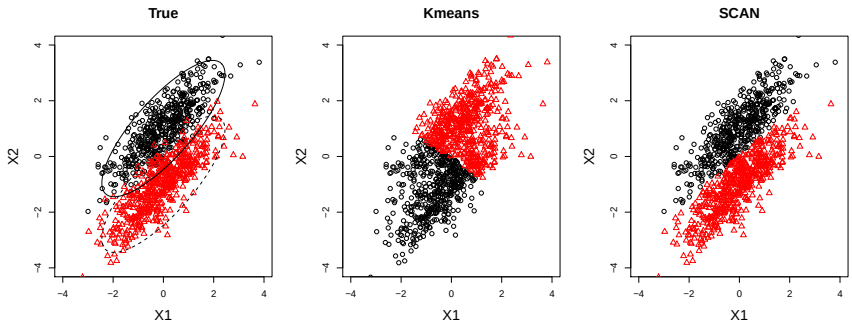

Fig. 2.1. The first plot represents the true clusters shown in red and black in the example of Section 2.3.2. The middle and right plots show the clusters obtained from the standard K-means clustering (Kmeans) and our SCAN method.

two stage method which applies joint graphical lasso (JGL) to the clusters obtained from the K-means clustering. This suggests that an accurate clustering is critical for the estimation performance of heterogeneous graphical models.

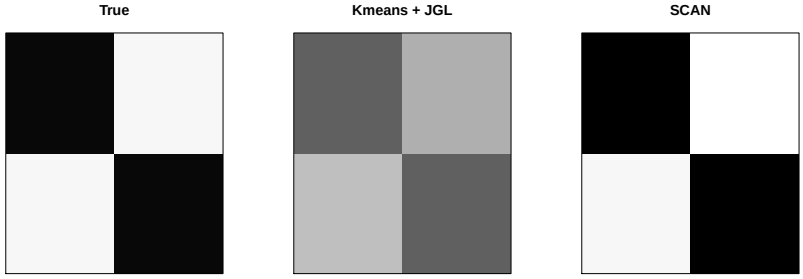

Fig. 2.2. The true precision matrix and the estimated precision matrices from the two stage method (Kmeans + JGL) and our SCAN method in the example of Section 2.3.2.

2.3.3 Effect of Sample Size and Dimension

We investigate the effect of sample size and dimension in terms of the estimation error and computational time. First, we empirically demonstrate the derived upper bound (2.2.15) for the estimation error by drawing the error pattern of our precision matrix estimator against sample size and dimension. The setting is the same as Section 2.3.2 except that we consider a tri-diagonal convariance structure. The results are summarized in Figure 2.3.3. In the first plot, we fix the dimension to be 10 and vary the sample size from 400 to 2000. In the second plot, we fix the sample size to be 5000 and vary the dimension from 5 to 50. The box plot refers to the the actual numerical values of precision matrix estimation errors, and the red dot is the theoretical error rate in each scenario. These results demonstrate that the empirical errors match very well with the theoretical error bound.

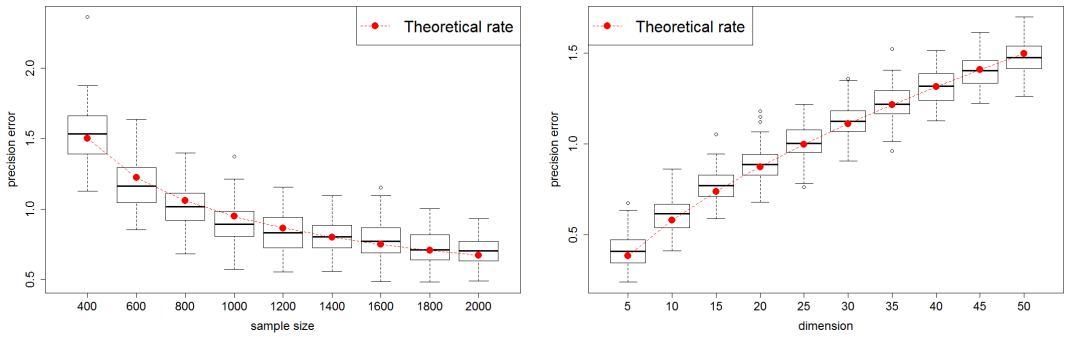

Fig. 2.3. Comparison of the numerical error and the theoretical error rates of our SCAN method. The left panel displays the precision matrix estimation error with varying sample sizes. The right panel displays the precision matrix estimation error with varying dimensions.

Second, we compare the average running time of our SCAN algorithm with varying sample sizes and dimensions. Figure 2.3.3 shows that our algorithm scales linearly with the sample size and roughly linearly with the dimension. This illustrates the efficiency and scalability of our proposed algorithm.

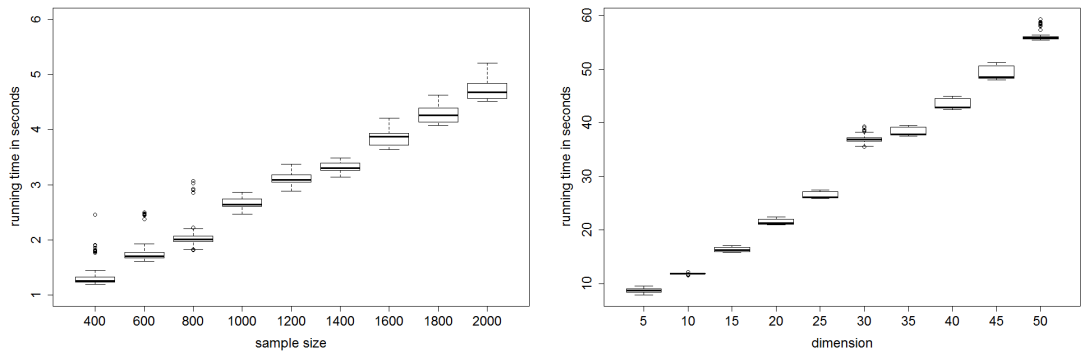

Fig. 2.4. Running time of our algorithm. The left panel is the running time with varying sample sizes and fixed dimension $p = 10$. The right panel is the running time with varying dimensions and fixed sample size $n = 5000$.

2.3.4 Simulations

In this subsection, we conduct extensive simulation studies to evaluate the performance of our algorithm. To assess the clustering performance of various methods, we compute the following clustering error (CE) which calculates the distance between an estimated clustering assignment $\widehat{\psi}$ and the true assignment ψ of the sample data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ (21; 75),

$$\mathrm{CE}(\widehat{\psi}, \psi) := \binom{n}{2}^{-1} \left| \{(i,j) : \mathbb{1}(\widehat{\psi}(\boldsymbol{X}_i) = \widehat{\psi}(\boldsymbol{X}_j)) \neq \mathbb{1}(\psi(\boldsymbol{X}_i) = \psi(\boldsymbol{X}_j)); i < j\} \right|,$$

where $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. To measure the estimation quality, we calculate the precision matrix error (PME) and cluster mean error (CME)

$$\mathrm{PME} := \frac{1}{K} \sum_{k=1}^{K} \left\| \widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}^{(k)} \right\|_F; \ \mathrm{CME} := \frac{1}{K} \sum_{k=1}^{K} \left\| \widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}^{(k)} \right\|_2.$$

Finally, to compare the variable selection performance, we compute the true positive rate (TPR, percentage of true edges selected) and the false positive rate (FPR, percentage of false edges selected)

$$\text{TPR} := \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i<j}\mathbb{1}(\omega_{kij}\neq 0,\widehat{\omega}_{kij}\neq 0)}{\sum_{i<j}\mathbb{1}(\omega_{kij}\neq 0)},$$
$$\text{FPR} := \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i<j}\mathbb{1}(\omega_{kij}= 0,\widehat{\omega}_{kij}\neq 0)}{\sum_{i<j}\mathbb{1}(\omega_{kij}= 0)}.$$

In the simulation, a three-class problem is considered. We illustrate three different types of network structures. In the first scenario, the network is assumed to have some regular structures. We generate a 5-block tridiagonal precision matrix with p features for the precision matrix. To allow the similarity of precision matrices across clusters, we set the off-diagonal entry of $\Omega_1, \Omega_2, \Omega_3$ as η, 0.99η, and 1.01η, respectively. The diagonal entries of Ω_1, Ω_2, and Ω_3 are all 1.

In the second and third scenarios, followed by (8), we simulate each network consisting of disjointed modules since many large networks in the real life exhibit a modular structure comprised of many disjointed or loosely connected components of relatively small size (76). Thus, each of three networks is generated with p features, which has ten equally sized unconnected subnetworks. Among the ten subnetworks, eight have the same structure and edge values across all the three classes, one remains the same only for the first two classes and the last one appears only for the first class. For the cluster structure of subnetwork, we consider two scenarios: power-law network and chain network, which are generated using the algorithm in (76) and (77). The detail construction is described as below.

Power-law network. Given an undirected network structure above, the initial ten-block precision matrix $(w_{ij}^1)_{p\times p}$ is generated by

$$w_{ij}^1 = \begin{cases} 1 & i\neq j; \\ 0 & i\neq j, \text{ no edge}; \\ \text{Unif}([-0.4,-0.1]\cup[0.1,0.4]) & i\neq j, \text{ edge exits}; \end{cases}$$

To ensure positive definiteness and symmetry, we divide each off-diagonal entry by 0.9 times the sum of the absolute values of off-diagonal entries in its row and average this rescaled matrix with its transpose. Denote the final transformed matrix by $\boldsymbol{A}$. The covariance matrix corresponding to the first class is created by

$$\boldsymbol{\Sigma}_{1ij} = d_{ij} \frac{\boldsymbol{A}_{ij}^{-1}}{\sqrt{\boldsymbol{A}_{ii}^{-1}\boldsymbol{A}_{jj}^{-1}}} \tag{2.3.2}$$

where $d_{ij} = 0.9$ for non-diagonal entry and $d_{ij} = 1$ for diagonal entry. For the covariance matrix corresponding to the second class, we create $\boldsymbol{\Sigma}_2$ be identical to $\boldsymbol{\Sigma}_1$ but reset one of ten block matrix to the identity matrix. Similarly, we reset one additional block matrix for $\boldsymbol{\Sigma}_3$.

Chain network. In the scenario, each of ten blocks of the first covariance matrix $\boldsymbol{\Sigma}_1$ is constructed in the following way. The ij-th element of each block has the form $\sigma_{ij} = \exp(-a|s_i - s_j|)$, where $s_1 < s_2 < \cdots < s_{p/10}$ for some $a > 0$. This is related to the autoregressive process of order one. In our case, we choose $a = 1$ and $s_i - s_{i-1} \sim \text{Unif}(0.5, 1)$ for $i = 2, \ldots, p/10$. Similarly, we create $\boldsymbol{\Sigma}_2$ be identical to $\boldsymbol{\Sigma}_1$ but reset one of ten block matrix to the identity matrix and reset one additional block matrix for $\boldsymbol{\Sigma}_3$.

After the networks are constructed, the samples are generated as follows. First, the cluster membership Y_i's are uniformly sampled from $\{1, 2, 3\}$. Given the cluster label, we generate each sample $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}(Y_i), \boldsymbol{\Sigma}(Y_i))$. Here, the cluster mean $\boldsymbol{\mu}(Y_i)$ is sparse, where its first 10 variables are of the form

$$(\mu \mathbf{1}_5^\top, -\mu \mathbf{1}_5^\top)^\top \mathbb{1}(Y_i = 1) + \mu \mathbf{1}_{10} \mathbb{1}(Y_i = 2) + (-\mu \mathbf{1}_5^\top, -\mu \mathbf{1}_5^\top)^\top \mathbb{1}(Y_i = 3),$$

with $\mathbf{1}_5$ being a 5-dimensional vector of all ones, and its last $p - 10$ variables are zeros. For the first scenario, we consider 3 simulation models with varying choices of μ and η:

- Model 1: $\mu = 0.8$ and $\eta = 0.3$,

- Model 2: $\mu = 1$ and $\eta = 0.3$,
- Model 3: $\mu = 1$ and $\eta = 0.4$.

Here μ controls the separability of the three clusters with larger μ corresponding to an easier clustering problem, and η represents the similarity level of precision matrices across clusters. For the second and third scenarios, we considered three simulation models with sequential choices of μ:

- Models 4,7: $\mu = 0.7$,
- Models 5,8: $\mu = 0.8$,
- Models 6,9: $\mu = 0.9$.

The number of features p is equal to 100 and sample size is equal to 300. The results are averaged over 50 experiments. The code is written in R and implemented on an Intel Xeon-E5 processor with 64 GB of RAM. The average computation time for SCAN of a single run took one and half minute.

In the experiment, our method selected the tuning parameters via the BIC criterion in Section 2.3.1. For a fair comparison, we also used the same tuning parameters λ_1, λ_2 in (22), and the same λ_2, λ_3 in the joint graphical lasso penalty of the two-stage approach. We repeated the procedure 50 times and reported the averaged clustering errors, estimation errors, and variable selection errors for each method as well as their standard errors. Table 2.2 is for regular network, Table 2.3 is for power-law networks and Table 2.4 is for chain networks. As shown in Table 2.3 and Table 2.4, the standard K-means clustering method has the largest clustering error due to a violation of its diagonal covariance matrix assumption. This will result in poor estimation for multiple precision matrices. The method of (22) improves the clustering performance of the standard K-means by using a graphical lasso in the precision matrix estimation. However, it obtains a relatively large precision matrix estimation error and very bad false positive rate since it ignores the similarity across different precision matrices. In contrast, our SCAN algorithm achieves the best

clustering accuracy and best precision matrix estimation accuracy for both scenarios. This is due to our simultaneous clustering and estimation strategy as well as the consideration of similarity of precision matrices across clusters. This experiment shows that a satisfactory clustering algorithm is critical to achieve accurate estimations of heterogeneous graphical models, and alternatively good estimation of the graphical model can also improve the clustering performance. This explains the success of our simultaneous method in terms of both clustering and graphical model estimation.

Table 2.2.
Simulation results of regular network. The clustering errors (CE), cluster mean errors (CME), precision matrix errors (PME), true positive rates (TPR) and false positive rates (FPR) of precision matrix estimation of four methods. The minimal clustering error and minimal estimation error in each simulation are shown in bold.

Models	Methods	CE	CME	PME	TPR /FPR
	K-means	$0.166_{0.011}$	$2.256_{0.108}$	NA	NA /NA
Model 1	K-means + JGL	$0.166_{0.011}$	$2.256_{0.108}$	$8.206_{0.090}$	$0.985_{0.001}$ /$0.023_{0.001}$
$\mu = 0.8$	(22)	$0.104_{0.007}$	$1.190_{0.052}$	$10.458_{0.0509}$	$0.960_{0.002}$ /$0.107_{0.001}$
$\eta = 0.3$	SCAN	$\mathbf{0.071_{0.007}}$	$\mathbf{1.120_{0.063}}$	$\mathbf{7.620_{0.072}}$	$\mathbf{0.993_{0.001}}$ /$\mathbf{0.022_{0.001}}$
	K-means	$0.210_{0.009}$	$3.428_{0.114}$	NA	NA/NA
Model 2	K-means + JGL	$0.210_{0.009}$	$3.428_{0.114}$	$12.099_{0.317}$	$0.989_{0.001}$ /$0.039_{0.003}$
$\mu = 1$	(22)	$0.125_{0.012}$	$1.860_{0.118}$	$12.833_{0.253}$	$0.993_{0.001}$ /$0.119_{0.006}$
$\eta = 0.3$	SCAN	$\mathbf{0.058_{0.012}}$	$\mathbf{1.476_{0.145}}$	$\mathbf{10.301_{0.332}}$	$\mathbf{0.997_{0.001}}$ /$\mathbf{0.036_{0.002}}$
	K-means	$0.021_{0.002}$	$1.289_{0.013}$	NA	NA /NA
Model 3	K-means + JGL	$0.021_{0.002}$	$1.289_{0.013}$	$7.639_{0.061}$	$0.993_{0.001}$ /$0.029_{0.002}$
$\mu = 1$	(22)	$0.021_{0.002}$	$0.968_{0.018}$	$10.115_{0.047}$	$0.968_{0.001}$ /$0.106_{0.001}$
$\eta = 0.4$	SCAN	$\mathbf{0.014_{0.001}}$	$\mathbf{0.956_{0.018}}$	$\mathbf{7.614_{0.061}}$	$\mathbf{0.993_{0.001}}$ /$\mathbf{0.029_{0.002}}$

Table 2.3.
Simulation results of power-law network. The clustering errors (CE), cluster mean errors (CME), precision matrix errors (PME), true positive rates (TPR) and false positive rates (FPR) of precision matrix estimation of four methods. The minimal clustering error and minimal estimation error in each simulation are shown in bold.

Models	Methods	CE	CME	PME	TPR /FPR
	K-means	$0.331_{0.007}$	$3.282_{0.047}$	NA	NA /NA
Model 4	K-means + JGL	$0.331_{0.007}$	$3.282_{0.047}$	$49.516_{0.159}$	$0.575_{0.002}$ /$0.034_{0.002}$
$\mu = 0.7$	(22)	$0.311_{0.006}$	$2.494_{0.055}$	$50.945_{0.164}$	$0.578_{0.002}$ /$0.134_{0.002}$
	SCAN	$\mathbf{0.283_{0.008}}$	$\mathbf{2.385_{0.065}}$	$\mathbf{48.845_{0.146}}$	$0.577_{0.003}$ /$0.032_{0.002}$
	K-means	$0.228_{0.010}$	$2.777_{0.111}$	NA	NA/NA
Model 5	K-means + JGL	$0.228_{0.010}$	$2.777_{0.111}$	$48.601_{0.132}$	$0.582_{0.002}$ /$0.044_{0.003}$
$\mu = 0.8$	(22)	$0.186_{0.011}$	$1.837_{0.113}$	$49.289_{0.122}$	$0.584_{0.001}$ /$0.131_{0.001}$
	SCAN	$\mathbf{0.156_{0.012}}$	$\mathbf{1.789_{0.119}}$	$\mathbf{47.729_{0.118}}$	$0.583_{0.002}$ /$0.041_{0.002}$
	K-means	$0.083_{0.010}$	$1.624_{0.120}$	NA	NA /NA
Model 6	K-means + JGL	$0.083_{0.010}$	$1.624_{0.120}$	$46.879_{0.093}$	$0.589_{0.002}$ /$0.070_{0.003}$
$\mu = 0.9$	(22)	$0.050_{0.002}$	$1.003_{0.018}$	$47.503_{0.003}$	$0.591_{0.001}$ /$0.128_{0.001}$
	SCAN	$\mathbf{0.045_{0.002}}$	$\mathbf{1.003_{0.018}}$	$\mathbf{46.356_{0.086}}$	$0.589_{0.001}$ /$0.068_{0.003}$

2.3.5 Glioblastoma Cancer Data Analysis

In this section, we apply our simultaneous clustering and graphical model estimation method to a Glioblastoma cancer dataset. We aim to cluster the glioblastoma multiforme (GBM) patients and construct the gene regulatory network of each subtype in order to improve our understanding of the GBM disease.

The raw gene expression dataset measures 17814 levels of mRNA expression of 482 GBM patients. Each patient belongs to one of four subgroups of GBM: Classical, Mesenchymal, Neural, and Proneural (7). Although they are biologically different,

Table 2.4.
Simulation results of chain network. The clustering errors (CE), cluster mean errors (CME), precision matrix errors (PME), true positive rates (TPR) and false positive rates (FPR) of precision matrix estimation of four methods. The minimal clustering error and minimal estimation error in each simulation are shown in bold.

Models	Methods	CE	CME	PME	TPR /FPR
	K-means	$0.277_{0.005}$	$2.705_{0.070}$	NA	NA /NA
Model 7	K-means + JGL	$0.277_{0.005}$	$2.705_{0.070}$	$25.608_{0.183}$	$0.995_{0.000}$ /$0.033_{0.001}$
$\mu = 0.7$	(22)	$0.267_{0.006}$	$1.815_{0.075}$	$29.341_{0.109}$	$0.991_{0.001}$ /$0.131_{0.002}$
	SCAN	$\mathbf{0.231_{0.007}}$	$\mathbf{1.652_{0.087}}$	$\mathbf{25.110_{0.106}}$	$0.991_{0.001}$ /$0.031_{0.001}$
	K-means	$0.200_{0.008}$	$2.124_{0.098}$	NA	NA/NA
Model 8	K-means + JGL	$0.200_{0.008}$	$2.124_{0.098}$	$24.499_{0.127}$	$0.996_{0.000}$ /$0.042_{0.001}$
$\mu = 0.8$	(22)	$0.168_{0.004}$	$1.055_{0.076}$	$27.494_{0.121}$	$0.995_{0.001}$ /$0.131_{0.001}$
	SCAN	$\mathbf{0.140_{0.004}}$	$\mathbf{1.046_{0.038}}$	$\mathbf{23.804_{0.085}}$	$0.996_{0.000}$ /$0.039_{0.001}$
	K-means	$0.123_{0.005}$	$1.465_{0.040}$	NA	NA /NA
Model 9	K-means + JGL	$0.123_{0.005}$	$1.465_{0.040}$	$23.663_{0.097}$	$0.997_{0.000}$ /$0.044_{0.001}$
$\mu = 0.9$	(22)	$0.116_{0.003}$	$1.031_{0.022}$	$26.476_{0.090}$	$0.996_{0.001}$ /$0.131_{0.001}$
	SCAN	$\mathbf{0.098_{0.003}}$	$\mathbf{1.025_{0.022}}$	$\mathbf{23.425_{0.083}}$	$0.998_{0.000}$ /$0.043_{0.002}$

these four subtypes share many similarities since they are all GBM diseases. For our analysis, we considered the 840 signature genes established by (7). Following the preprocess procedures in (9), we excluded the genes with no subtype information or the genes with missing values. We then applied the sure independence screening analysis (78) to finally include 50 genes in our analysis. These 50 signature genes are highly distinctive for these four subtypes. In the analysis, we pretended that the subtype information of each patient was unknown and evaluated the clustering accuracy of various clustering methods by comparing the estimated groups with the true subtypes. In all methods, we fixed $K = 4$. Moreover, we set the tuning parameters

$\lambda_1 = 0.065, \lambda_2 = 0.238$, and $\lambda_3 = 0.138$ in our SCAN algorithm. For a fair comparison, we also used the same λ_1, λ_2 in (22), and the same λ_2, λ_3 in the joint graphical lasso of the two-stage method.

Table 2.5 reported the clustering errors of all methods as well as the number of informative variables in the corresponding estimated means and precision matrices. The standard K-means clustering has the large clustering error due to its ignorance of the network structure in the precision matrices. Therefore, the consequent joint graphical lasso method of the network reconstruction is less reliable. The method in (22) performed even worse. This is because their method estimates each precision matrix individually without borrowing information from each other. In this gene network example, all of the four graphical models share many edges due to the commonality in the GBM diseases. (22)'s method may suffer from the small sample size. Our method is able to achieve the best clustering performance due to the procedure of simultaneous clustering and heterogeneous graphical model estimation.

Table 2.5.
The clustering errors and the number of selected features in cluster mean and precision matrix of various methods in the Glioblastoma Cancer Data.

Methods	Clustering Error	$\sum_k \|\widehat{\boldsymbol{\mu}}^{(k)}\|_0$	$\sum_k \|\widehat{\boldsymbol{\Omega}}^{(k)}\|_0$
K-means	0.262	200	NA
(22)	0.336	106	1820
K-means + JGL	0.262	200	1360
SCAN	**0.222**	128	1452

To evaluate the ability of reconstructing gene regulatory network of each subtype, we report the four gene networks estimated from our SCAN method in Figure 1.1. The black lines are links shared in all subtypes, and the color lines are uniquely presented in some subtypes. Clearly, most edges are black lines, which indicates the common structure of all subtypes. For instance, the link between ZNF45 and

ZNF134 is significant across all the four subtypes. Those two genes belong to ZNF gene family. They are known to play roles in making zinc finger proteins, which are regulatory proteins that are functional important to many cellulars. As they play roles in the same biological process, it is reasonable to expect this link is shared by all GBM subtypes. There are two links that shared by three subtypes except neural subtype: TNFRSF1B↔TRPM2, PTPRC↔ TRPM2. One link uniquely appears in Proneural subtype: ACTR1A ↔DWED and one link FBXO3↔HMG20B is uniquely shown in neural subtype. These findings agree with the existing results in (7). It has been shown that the PTPRC is a well-described microglia marker and is highly exposed in the set of murine astrocytic samples which are strongly associated with the Mesenchymal group. In addition, TRPM2 and TNFRSF1B are shown frequently in the GOTERM category of Mesenchymal group but less likely to appear in Neural group. And FBXO3 is only significant in the cell part of neural subtype. Furthermore, ACTR1A is only found in the intracellular non-membrane-bound organelle and protein binding of Proneural subtype in the supplemental material of (7). It would also be of interest to investigate unique gene links that were not discovered in existing literatures for better understanding of GBM diseases.

2.4 Discussion

In this paper, we propose a new SCAN method for simultaneous clustering and estimation of heterogeneous graphical models with common structures. We describe the theoretical properties of SCAN and we show that the estimation error bound of our SCAN algorithm consists of statistical error and optimization error, which explicitly addresses the trade-off between statistical accuracy and computational complexity. In our experiments, the tuning parameters can be chosen via an efficient BIC-type criterion. For future work, it is of interest to investigate the model selection consistency of these tuning parameters and study the distributed implementation of ECM algorithm based on the work in (79).

2.5 Main Proofs

In this section, we provide detailed proofs of key results: Theorem 1 and Corollary 2.2.12. The proofs of other lemmas and theorems are deferred to the next section.

2.5.1 Proof of Theorem 1

First we introduce some notation. Recall the definition of support space $\mathcal{M}$ in (2.2.2). The orthogonal complement of support space $\mathcal{M}$, namely, is defined as the set

$$\mathcal{M}^{\perp} := \{\boldsymbol{\Theta}' \in \Xi \mid \langle \boldsymbol{V}, \boldsymbol{\Theta}' \rangle = 0 \text{ for all } \boldsymbol{V} \in \mathcal{M}\}.$$

The projection operator $\Pi_{\mathcal{M}}(\boldsymbol{\Theta}) : \Xi \to \Xi$ is defined as

$$\Pi_{\mathcal{M}}(\boldsymbol{\Theta}) := \arg\min_{\boldsymbol{V} \in \mathcal{M}} \|\boldsymbol{V} - \boldsymbol{\Theta}\|_2 .$$

To simplify the notation, we frequently use the shorthand $\boldsymbol{\Theta}_{\mathcal{M}} = \Pi_{\mathcal{M}}(\boldsymbol{\Theta})$ and $\boldsymbol{\Theta}_{\mathcal{M}^{\perp}} = \Pi_{\mathcal{M}^{\perp}}(\boldsymbol{\Theta})$.

In order to efficiently solve the high-dimensional regularized problem, we explore some good properties enjoyed by SCAN penalty in Lemma 2.5.1 and Lemma 2.5.2. Similar properties can be derived by any decomposable penalty.

Lemma 2.5.1. The SCAN penalty $\mathcal{P}$ is convex and decomposable with respect to $(\mathcal{M}, \mathcal{M}^{\perp})$. In detail,

$$\mathcal{P}(\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2) = \mathcal{P}(\boldsymbol{\Theta}_1) + \mathcal{P}(\boldsymbol{\Theta}_2), \text{ for any } \boldsymbol{\Theta}_1 \in \mathcal{M}, \boldsymbol{\Theta}_2 \in \mathcal{M}^{\perp}.$$

The dual norm of SCAN penalty $\mathcal{P}$ is given by

$$\mathcal{P}^*(\boldsymbol{\Theta}) := \max_{i,j,k,i \neq j} \left(M_1 \sqrt{\mu_{kj}^2}, M_2 \sqrt{\omega_{kij}^2}, M_3 \left(\sum_{k=1}^{K} \omega_{kij}^2 \right)^{1/2} \right). \tag{2.5.1}$$

Proof of Lemma 2.5.1: The convexity of SCAN comes from the convexity of lasso penalty for cluster means and the convexity of group graphical lasso penalty for precision matrices. The decomposability and derivation of dual norm is obvious from the definition. Also see (70). ■

Lemma 2.5.2. For all vectors $\boldsymbol{\Theta}$ belonging to support space $\mathcal{M}$, $\mathcal{P}(\boldsymbol{\Theta}_{\mathcal{M}})$ satisfies the following inequality:

$$\mathcal{P}(\boldsymbol{\Theta}_{\mathcal{M}}) \leq \nu(\mathcal{M}) \left\|\boldsymbol{\Theta}_{\mathcal{M}}\right\|_2, \tag{2.5.2}$$

where $\nu(\mathcal{M}) = M_1\sqrt{Kd} + (M_2\sqrt{K} + M_3)\sqrt{s}$ is the *support space compatibility constant* defined in (2.2.12) .

Proof of Lemma 2.5.2: The detailed proof of Lemma 2.5.2 is discussed in 2.6.5. ■

Next lemma is a key step to establish our main theorem. It quantifies the estimation error in one iteration step. According to this lemma, one can precisely understand how the statistical error and optimization error accumulate with more and more iterations.

Lemma 2.5.3. Suppose $\boldsymbol{\Theta}^*$ lies in the interior of Ξ. If $\boldsymbol{\Theta}^{(t-1)} \in \mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$, with choice of $\lambda_n^{(t)} = \varepsilon + \tau\|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2/\nu(\mathcal{M})$, final estimation error satisfies $\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\|_2 \leq 6\nu(\mathcal{M})\lambda_n^{(t)}/\gamma$ with probability at least $1 - \delta'$ for all $t = 1, 2, \ldots$. Here τ, λ and $\nu(\mathcal{M})$ are defined in Lemma 2.2.3, Lemma 2.2.5 and Lemma 2.5.2 accordingly.

Proof of Lemma 2.5.3: Proof is postponed to section 2.5.3. ■

Equipped with Lemmas 2.5.3, we are able to precisely quantify the final estimation error after t iteration steps. This can be achieved by mathematical induction. For simplicity, define $\kappa := 6\tau/\gamma$. When $t = 1$, we have $\boldsymbol{\Theta}^{(0)} \in \mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$. Applying Lemma 2.5.3 yields that

$$\begin{aligned} \left\|\boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^*\right\|_2 &\leq \frac{6\lambda_n^{(1)}\nu(\mathcal{M})}{\gamma} \\ &= \frac{6\nu(\mathcal{M})}{\gamma}\varepsilon + \kappa\left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2. \end{aligned}$$

Suppose the following inequality is true for some $t \geq 1$,

$$\left\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\right\|_2 \leq \frac{1-\kappa^t}{1-\kappa}\frac{6\nu(\mathcal{M})}{\gamma}\varepsilon + \kappa^t\left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2,$$

with probability at least $1-t\delta'$. We need to verify when $t=t+1$, the above inequality still holds. First, we show that $\boldsymbol{\Theta}^{(t)}$ is within a ball of $\boldsymbol{\Theta}^*$ with radius α. Under the assumption that $\varepsilon \leq (1-\kappa)\alpha\gamma/(6\nu(\mathcal{M}))$ for sufficient large n, we have

$$\begin{aligned}\left\|\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Theta}^*\right\|_2 &\leq \frac{1-\kappa^t}{1-\kappa}\frac{6\nu(\mathcal{M})}{\gamma}\frac{(1-\kappa)\alpha\gamma}{6\nu(\mathcal{M})}+\kappa^t\left\|\boldsymbol{\Theta}^{(0)}-\boldsymbol{\Theta}^*\right\|_2 \\ &\leq (1-\kappa^t)\alpha+\kappa^t\alpha=\alpha.\end{aligned}$$

Consequently, we have $\boldsymbol{\Theta}^{(t)}\in\mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$. Applying Lemma 2.5.3 with $t+1$ implies that

$$\begin{aligned}\left\|\boldsymbol{\Theta}^{(t+1)}-\boldsymbol{\Theta}^*\right\|_2 &\leq \frac{6\nu(\mathcal{M})\varepsilon}{\gamma}+\kappa\left\|\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Theta}^*\right\|_2 \\ &\leq \frac{6\nu(\mathcal{M})\varepsilon}{\gamma}+\kappa\left(\frac{1-\kappa^t}{1-\kappa}\frac{6\nu(\mathcal{M})\varepsilon}{\gamma}+\kappa^t\left\|\boldsymbol{\Theta}^{(0)}-\boldsymbol{\Theta}^*\right\|_2\right) \\ &= \frac{1-\kappa^{t+1}}{1-\kappa}\frac{6\nu(\mathcal{M})\varepsilon}{\gamma}+\kappa^{t+1}\left\|\boldsymbol{\Theta}^{(0)}-\boldsymbol{\Theta}^*\right\|_2,\end{aligned}$$

with probability at least $1-(t+1)\delta'$. Therefore, we reach the conclusion that

$$\begin{aligned}\left\|\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Theta}^*\right\|_2 &\leq \frac{1-\kappa^t}{1-\kappa}\frac{6\nu(\mathcal{M})}{\gamma}\varepsilon+\kappa^t\left\|\boldsymbol{\Theta}^{(0)}-\boldsymbol{\Theta}^*\right\|_2 \\ &\leq \frac{6\nu(\mathcal{M})\varepsilon}{(1-\kappa)\gamma}+\kappa^t\left\|\boldsymbol{\Theta}^{(0)}-\boldsymbol{\Theta}^*\right\|_2,\end{aligned}$$

with probability at least $1-t\delta'$. This concludes the proof of Theorem 1. ■

2.5.2 Proof of Corollary 2.2.12

It is worth to notice that sufficiently large iterations ensure that the optimization error will be dominated by statistical error finally as $\kappa<1/2$. First we provide a stopping rule T. Plugging $\varepsilon_1,\varepsilon_2$ from (2.6.14) & (2.6.15) into statistical error part and letting $\delta=1/p$, we have:

$$\begin{aligned}SE=&\frac{1}{1-\kappa}\frac{6}{\gamma}\left[\left(\sqrt{Kd}+(\sqrt{K}+1)\sqrt{s}\right)\left(CK\|\boldsymbol{\Omega}^*\|_\infty+C'K^{1.5}\right)\sqrt{\frac{\log p}{n}}\right] \\ &+\frac{1}{1-\kappa}\frac{6}{\gamma}\left[C''\sqrt{p}\sqrt{\frac{K^3\log p}{n}}\right].\end{aligned}$$

Note that under Condition 2.2.11, $K = o(p)$. Then SE is simplified by

$$SE \leq \frac{6\widetilde{C}}{(1-\kappa)\gamma}\|\mathbf{\Omega}^*\|_\infty \left(\sqrt{Kd} + \sqrt{Ks+p}\right)\sqrt{\frac{K^3 \log p}{n}},$$

for some constant $\widetilde{C}$. For simplicity, let's denote

$$\varphi(n,p,K) = \frac{6\widetilde{C}}{(1-\kappa)\gamma}\|\mathbf{\Omega}^*\|_\infty \left(\sqrt{Kd} + \sqrt{Ks+p}\right)\sqrt{\frac{K^3 \log p}{n}}.$$

Therefore, the bound (2.2.14) suggests a reasonable choice of the number of iterations. In particular, when

$$t \geq T = \log_{1/\kappa}\left(\frac{\left\|\mathbf{\Theta}^{(0)} - \mathbf{\Theta}^*\right\|_2}{\varphi(n,p,K)}\right), \tag{2.5.3}$$

the optimization error is dominated by statistical error. Final estimation error will be upper bounded by

$$\left\|\mathbf{\Theta}^{(T)} - \mathbf{\Theta}^*\right\|_2 \leq \frac{12\widetilde{C}}{(1-\kappa)\gamma}\left(\|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^5 d \log p}{n}} + \|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^3(Ks+p)\log p}{n}}\right),$$

with probability at least $1 - T(26K^2 + 8K + 1)/p$. Plugging in the expression of T in (2.5.3), the probability term is bounded by:

$$\begin{aligned}\frac{T(26K^2+8K+1)}{p} &\lesssim \frac{\log_{1/\kappa}\left(n/\left(\left(\sqrt{Kd}+\sqrt{Ks+p}\right)\sqrt{K^3\log p}\right)\right)K^2}{p} \\ &\lesssim \frac{K^2 \log_{1/\kappa} n}{p}.\end{aligned}$$

Under Condition 2.2.11, $T(26K^2 + 8K + 1)/p$ goes to zero as K and p diverging. Putting pieces together, we have

$$\left\|\mathbf{\Theta}^{(T)} - \mathbf{\Theta}^*\right\|_2 \leq \frac{12\widetilde{C}}{(1-\kappa)\gamma}\left(\|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^5 d \log p}{n}} + \|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^3(Ks+p)\log p}{n}}\right),$$

which implies

$$\begin{aligned}&\sum_{k=1}^{K}\left(\left\|\boldsymbol{\mu}_k^{(T)} - \boldsymbol{\mu}_k^*\right\|_2 + \left\|\mathbf{\Omega}_k^{(T)} - \mathbf{\Omega}_k^*\right\|_F\right) \\ &\leq \frac{12\widetilde{C}}{(1-\kappa)\gamma}\left(\|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^5 d \log p}{n}} + \|\mathbf{\Omega}^*\|_\infty\sqrt{\frac{K^3(Ks+p)\log p}{n}}\right),\end{aligned}$$

with probability converging to 1. It ends the proof of Corollary 2.2.12. ■

2.5.3 Proof of Lemma 2.5.3

We first consider an unsymmetrized version of $\boldsymbol{\Theta}^{(t)}$. Our proof makes use of the function $f : \Xi \to \mathbb{R}$ given by:

$$f(\Delta) := Q_n(\boldsymbol{\Theta}^* + \Delta | \boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \lambda_n^{(t)} \left(\mathcal{P}(\boldsymbol{\Theta}^* + \Delta) - \mathcal{P}(\boldsymbol{\Theta}^*) \right).$$

This function helps us evaluate the error between the iterative estimator $\boldsymbol{\Theta}^{(t)}$ and the true parameter $\boldsymbol{\Theta}^*$. In addition, we exploit the following fact:

$$\begin{cases} f(0) = 0 \\ f(\widehat{\Delta}) \geq 0 \text{ when } \widehat{\Delta} = \boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*. \end{cases} \tag{2.5.4}$$

The second property is from the optimality of $\boldsymbol{\Theta}^{(t)}$ in terms of the sample version objective function. In detail,

$$\boldsymbol{\Theta}^{(t)} = \arg\max_{\boldsymbol{\Theta}'} Q_n(\boldsymbol{\Theta}' | \boldsymbol{\Theta}^{(t-1)}) - \lambda_n^{(t)} \mathcal{P}(\boldsymbol{\Theta}'). \tag{2.5.5}$$

Correspondingly, there is a classical result named self-consistency property for population version objective function in (80), which in detail is

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}'} Q(\boldsymbol{\Theta}' | \boldsymbol{\Theta}^*). \tag{2.5.6}$$

The whole proof follows two steps. In Step I, we show that $f(\Delta) < 0$ if $\|\Delta\|_2 = \xi$. Next in Step II, we show that the error term $\widehat{\Delta}$ must satisfy $\|\widehat{\Delta}\|_2 < \xi$ under the result in Step I.

Step I: we begin to establish an upper bound on $f(\Delta)$ over the set $\mathbb{C}(\xi) := \{\Delta : \|\Delta\|_2 = \xi\}$ for the chosen radius $\xi = 6\lambda_n^{(t)} \nu(\mathcal{M})/\gamma$. From the assumption on n, when n is large enough,

$$\varepsilon \leq \frac{(1-\kappa)\alpha\gamma}{6\nu(\mathcal{M})} \leq \frac{(2-\kappa)\alpha\gamma}{6\nu(\mathcal{M})},$$
$$\frac{6\nu(\mathcal{M})\varepsilon}{\gamma} \leq (2-\kappa)\alpha.$$

On the other hand, as $\|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2 \leq \alpha$, ξ satisfies,

$$\xi = \frac{6\nu(\mathcal{M})\varepsilon}{\gamma} + \kappa \left\| \boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^* \right\|_2 \leq 2\alpha.$$

It is sufficient to show that $\mathbb{C}(\xi) \subseteq \mathbb{C} = \{\Delta | \|\Delta\|_2 \leq 2\alpha\}$. According to Lemma 2.2.5, replacing $\boldsymbol{\Theta}' - \boldsymbol{\Theta}^*$ by Δ, then any $\Delta \in \mathbb{C}(\xi)$ enjoys restricted strong concavity property, which implies:

$$Q_n(\boldsymbol{\Theta}^* + \Delta | \boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) \leq \big\langle \nabla Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}), \Delta \big\rangle - \frac{\gamma}{2} \|\Delta\|_2^2,$$

with probability at least $1 - \delta$. Subtracting $\lambda_n^{(t)}(\mathcal{P}(\boldsymbol{\Theta}^* + \Delta) - \mathcal{P}(\boldsymbol{\Theta}^*))$ from both sides, we construct an upper bound of $f(\Delta)$ in the right side,

$$f(\Delta) \leq \underbrace{\big\langle \nabla Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}), \Delta \big\rangle}_{(i)} - \lambda_n^{(t)} \underbrace{\big(\mathcal{P}(\boldsymbol{\Theta}^* + \Delta) - \mathcal{P}(\boldsymbol{\Theta}^*)\big)}_{(ii)} - \frac{\gamma}{2} \|\Delta\|_2^2.$$

Bounding (i)**:** Note that Q_n is a sample version Q-function but $\boldsymbol{\Theta}^*$ comes from population version Q-function (2.5.6). So we use $\nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})$ as a bridge to connect the sample-based analysis and population-based analysis together.

$$\begin{aligned}
(i) &\leq \big| \big\langle \nabla Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) \\
&\qquad + \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^*), \Delta \big\rangle \big| \\
&\leq \underbrace{\big| \big\langle \nabla Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}), \Delta \big\rangle \big|}_{\text{Statistical Error(SE)}} \\
&\qquad + \underbrace{\big| \big\langle \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^*), \Delta \big\rangle \big|}_{\text{Optimization Error(OE)}}.
\end{aligned}$$

Note that $\boldsymbol{\Theta}^*$ lies in the interior of Ξ. According to the self-consistency property (2.5.6), $\nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^*) = 0$ which implies the first inequality holds. This decomposition for (i) leads to the optimization error part and statistical error part.

For simplicity, we write $h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) = \nabla Q_n(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})$. Since the group graphical lasso penalty does not penalize the diagonal element, it is a semi-norm. Recall that both Δ and $h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})$ are $K(p^2 + p)$ dimensional vectors. Then by the definition of $\mathcal{G}$ and $\mathcal{G}^c$ in (2.2.1), statistical error can be decomposed further by:

$$\begin{aligned}
\text{SE} &\leq \big| \langle h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})_{\mathcal{G}^c}, \Delta_{\mathcal{G}^c} \rangle \big| + \big| \langle h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})_{\mathcal{G}}, \Delta_{\mathcal{G}} \rangle \big| \\
&\leq \big\| h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})_{\mathcal{G}^c} \big\|_{\mathcal{P}^*} \cdot \mathcal{P}(\Delta_{\mathcal{G}^c}) + \| h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})_{\mathcal{G}} \|_2 \cdot \|\Delta_{\mathcal{G}}\|_2 \\
&\leq \| h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) \|_{\mathcal{P}^*} \cdot \mathcal{P}(\Delta) + \| h(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)})_{\mathcal{G}} \|_2 \cdot \|\Delta\|_2.
\end{aligned}$$

The second inequality comes from the generalized Cauchy-Schwarz inequality. After excluding the diagonal terms from precision matrices, $\mathcal{P}(\Delta_{\mathcal{G}^c})$ can be treated as a norm. The last inequality is because both the penalties $\mathcal{P}$ and $\mathcal{P}^*$ do not penalize the diagonal term of precision matrices. Under statistical error Condition 2.2.6,

$$\text{SE} \leq \varepsilon_1 \mathcal{P}(\Delta) + \varepsilon_2 \|\Delta\|_2, \tag{2.5.7}$$

with probability at least $1 - (\delta_1 + \delta_2)$.

On the other hand, from the assumption that $\boldsymbol{\Theta}^{(t-1)}$ is in the $\mathcal{B}_\alpha(\boldsymbol{\Theta}^*)$, we are able to apply the Gradient Stability condition in Lemma 2.2.3 to bound OE.

$$\begin{aligned} \text{OE} &\leq \|\nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^{(t-1)}) - \nabla Q(\boldsymbol{\Theta}^* | \boldsymbol{\Theta}^*)\|_2 \cdot \|\Delta\|_2 \\ &\leq \tau \|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2 \cdot \|\Delta\|_2. \end{aligned} \tag{2.5.8}$$

Therefore, putting (2.5.7) and (2.5.8) together, (i) is upper bounded by

$$(i) \leq \varepsilon_1 \mathcal{P}(\Delta) + \varepsilon_2 \|\Delta\|_2 + \tau \|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2 \cdot \|\Delta\|_2, \tag{2.5.9}$$

with probability at least $1 - (\delta_1 + \delta_2)$.

Bounding (ii)**:** The decomposability of SCAN penalty in Lemma 2.5.1 implies $\mathcal{P}(\boldsymbol{\Theta}^* + \Delta) = \mathcal{P}(\boldsymbol{\Theta}^* + \Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp})$. By triangle inequality, it is sufficient to bound (ii),

$$\begin{aligned} (ii) &= \mathcal{P}(\boldsymbol{\Theta}^* + \Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\boldsymbol{\Theta}^*) \\ &\geq \mathcal{P}(\boldsymbol{\Theta}^*) - \mathcal{P}(\Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\boldsymbol{\Theta}^*) \\ &= \mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\Delta_{\mathcal{M}}). \end{aligned} \tag{2.5.10}$$

Combining (2.5.9) and (2.5.10), $f(\Delta)$ is upper bounded by:

$$\begin{aligned} f(\Delta) \leq \varepsilon_1 \mathcal{P}(\Delta) + \varepsilon_2 \|\Delta\|_2 + \tau \|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2 \cdot \|\Delta\|_2 \\ - \lambda_n^{(t)} \left(\mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\Delta_{\mathcal{M}})\right) - \frac{\gamma}{2} \|\Delta\|_2^2. \end{aligned}$$

Triangle inequality implies $\mathcal{P}(\Delta) \leq \mathcal{P}(\Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp})$. After combining some terms, the right hand side above could be further bounded by:

$$\begin{aligned} f(\Delta) \leq -\frac{\gamma}{2} \|\Delta\|_2^2 + (\lambda_n^{(t)} + \varepsilon_1) \mathcal{P}(\Delta_{\mathcal{M}}) + (\varepsilon_1 - \lambda_n^{(t)}) \mathcal{P}(\Delta_{\mathcal{M}^\perp}) \\ + \varepsilon_2 \|\Delta\|_2 + \tau \|\boldsymbol{\Theta}^{(t-1)} - \boldsymbol{\Theta}^*\|_2 \cdot \|\Delta\|_2, \end{aligned} \tag{2.5.11}$$

with probability at least $1-(\delta+\delta_1+\delta_2)$. Let $\delta'=\delta+\delta_1+\delta_2$. According to Lemma 2.5.2, we have the inequality $\mathcal{P}(\Delta_{\mathcal{M}})\leq\nu(\mathcal{M})\|\Delta_{\mathcal{M}}\|_2$. By the definition of $\Pi_{\mathcal{M}}(\Delta)$, we have

$$\|\Delta_{\mathcal{M}}\|_2=\|\Pi_{\mathcal{M}}(\Delta)-\Pi_{\mathcal{M}}(0)\|_2\leq\|\Delta-0\|_2=\|\Delta\|_2.$$

Then $\mathcal{P}(\Delta_{\mathcal{M}})$ is further bounded by

$$\mathcal{P}(\Delta_{\mathcal{M}})\leq\nu(\mathcal{M})\|\Delta\|_2. \tag{2.5.12}$$

Substituting (2.5.12) into (2.5.11), we obtain:

$$\begin{aligned}f(\Delta)\leq&\left(\varepsilon_1+\frac{\varepsilon_2+\tau\|\boldsymbol{\Theta}^{(t-1)}-\boldsymbol{\Theta}^*\|_2}{\nu(\mathcal{M})}\right)\nu(\mathcal{M})\|\Delta\|_2-\frac{\gamma}{2}\|\Delta\|_2^2\\&+\lambda_n^{(t)}\nu(\mathcal{M})\|\Delta\|_2+(\varepsilon_1-\lambda_n^{(t)})\mathcal{P}(\Delta_{\mathcal{M}^\perp}),\end{aligned}$$

with at least probability $1-\delta'$. Recall that we choose

$$\lambda_n^{(t)}=\varepsilon+\frac{\tau\|\boldsymbol{\Theta}^{(t-1)}-\boldsymbol{\Theta}^*\|_2}{\nu(\mathcal{M})},\epsilon=\epsilon_1+\frac{\epsilon_2}{\nu(\mathcal{M})}.$$

From the construction of $\lambda_n^{(t)}$, the inequality $\varepsilon_1-\lambda_n^{(t)}<0$ always holds. Therefore, the upper bound for $f(\Delta)$ can be simplified by

$$\begin{aligned}f(\Delta)&\leq-\frac{\gamma}{2}\|\Delta\|_2^2+2\lambda_n^{(t)}\nu(\mathcal{M})\|\Delta\|_2\\&=-\frac{6(\lambda_n^{(t)}\nu(\mathcal{M}))^2}{\gamma}<0.\end{aligned}$$

where the above equality is due to $\Delta\in\mathbb{C}(\xi)$. Now we reach the conclusion that $f(\Delta)<0$ for all vectors $\Delta\in\mathbb{C}(\xi)$.

Step II: Now we start to prove the following statement: if for some optimal solution $\boldsymbol{\Theta}^{(t)}$ in (2.5.5), the corresponding error term $\widehat{\Delta}=\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Theta}^*$ satisfies the inequality $\|\widehat{\Delta}\|_2>\xi$, there must exist some vectors $\widetilde{\Delta}$ which belong to $\mathbb{C}(\xi)$ such that $f(\widetilde{\Delta})\geq 0$. Before our forward proofs, let's state a lemma which describe the curvature of function $Q_n(\cdot|\boldsymbol{\Theta}^{(t-1)})$.

Lemma 2.5.4. $Q_n(\cdot|\boldsymbol{\Theta}^{(t-1)})$ satisfies the following inequality a.s.:

$$Q_n(\boldsymbol{\Theta}^{(1)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n\left(\boldsymbol{\Theta}^{(2)}|\boldsymbol{\Theta}^{(t-1)}\right) \leq \left\langle \nabla Q_n\left(\boldsymbol{\Theta}^{(2)}\middle|\boldsymbol{\Theta}^{(t-1)}\right), \boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^{(2)} \right\rangle.$$

when $(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}) = (\boldsymbol{\Theta}^{(t)}, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$ or $(\boldsymbol{\Theta}^*, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$.

Proof of Lemma 2.5.4: The detailed proof of Lemma 2.5.4 is discussed in 2.6.6. ■

The lemma tells us that we only require sample-based Q-function to be point-wise concave rather than global concave. If $\|\widehat{\Delta}\|_2 > \xi$, then the line joining $\widehat{\Delta}$ to 0 must intersect the set $\mathbb{C}(\xi)$ at some intermediate points $t^*\widehat{\Delta}$, for some $t^* \in (0,1)$. According to Lemma 2.5.4,

$$\begin{aligned}
&Q_n(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}) \\
&\qquad \leq \left\langle \nabla Q_n(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}), (1-t^*)(\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*) \right\rangle \\
&Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}) - Q_n\left(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}\right) \\
&\qquad \leq \left\langle \nabla Q_n\left(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}\right), -t^*(\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*) \right\rangle.
\end{aligned}$$

Adding the above two inequalities together with proper scaling, we can get

$$t^*Q_n(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t-1)}) + (1-t^*)Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}) \leq Q_n(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}).$$

According to the convexity of $\mathcal{P}(\boldsymbol{\Theta})$,

$$\begin{aligned}
&\mathcal{P}\left(\boldsymbol{\Theta}^* + t^*\widehat{\Delta}\right) - \mathcal{P}\left(\boldsymbol{\Theta}^*\right) = \mathcal{P}\left(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*\right) - \mathcal{P}(\boldsymbol{\Theta}^*) \\
&\leq t^*\mathcal{P}(\boldsymbol{\Theta}^{(t)}) + (1-t^*)\mathcal{P}(\boldsymbol{\Theta}^*) - \mathcal{P}(\boldsymbol{\Theta}^*) = t^*\left(\mathcal{P}(\boldsymbol{\Theta}^{(t)}) - \mathcal{P}(\boldsymbol{\Theta}^*)\right).
\end{aligned}$$

Putting the above pieces together, it is shown that

$$\begin{aligned}
f(t^*\widehat{\Delta}) \quad &= Q_n\left(t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}\right) - Q_n\left(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}\right) \\
&\qquad - \lambda_n^{(t)}\left(\mathcal{P}(\boldsymbol{\Theta}^* + \Delta) - \mathcal{P}(\boldsymbol{\Theta}^*)\right) \\
&\geq t^*\left(Q_n(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^{(t-1)}) - \lambda_n^{(t)}\left(\mathcal{P}(\boldsymbol{\Theta}^{(t)}) - \mathcal{P}(\boldsymbol{\Theta}^*)\right)\right) \\
&= t^*f(\widehat{\Delta}).
\end{aligned}$$

On the other hand, the optimality property (2.5.4) guarantees $f(\widehat{\Delta}) \geq 0$, and hence $f(t^*\widehat{\Delta}) \geq 0$ as well. Thus, we have constructed a vector $\widetilde{\Delta} = t^*\widehat{\Delta}$ with the claimed properties. This proves the statement in the beginning of Step II. Therefore, combining with the result in Step I, the contradiction of the statement in Step II implies that

$$\left\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\right\|_2 \leq \xi = \frac{6\lambda_n^{(t)}\nu(\mathcal{M})}{\gamma}, \tag{2.5.13}$$

with probability at least $1 - \delta'$. This concludes the proof of Lemma 2.5.3. ■

2.6 Additional Proofs

2.6.1 Proof of Lemma 2.2.1

The result follows by setting the derivative of $Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta})$ with respect to $\boldsymbol{\mu}_k^{'}$ or $\boldsymbol{\Omega}_k^{'}$ as zero. In particular, solving

$$\frac{\partial Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta})}{\partial \boldsymbol{\mu}_k^{'}} = \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\Omega}_k^{'}(\boldsymbol{X} - \boldsymbol{\mu}_k^{'})] = 0,$$

implies that

$$\arg\max_{\boldsymbol{\mu}_k^{'}} Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta}) = \frac{[\boldsymbol{\Omega}_k^{'}]^{-1}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\Omega}_k^{'}\boldsymbol{X}]}{\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]} = \frac{\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}]}{\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]}.$$

Similarly, solving

$$\frac{\partial Q(\boldsymbol{\Theta}^{'}|\boldsymbol{\Theta})}{\partial \boldsymbol{\Omega}_k^{'}} = \frac{1}{2}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})][\boldsymbol{\Omega}_k^{'}]^{-1} - \frac{1}{2}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{\mu}_k^{'})(\boldsymbol{X} - \boldsymbol{\mu}_k^{'})^{\top}] = 0,$$

implies (2.2.6). This ends the proof of Lemma 2.2.1. ■

2.6.2 Proof of Lemma 2.2.3

We consider k-th group first

$$\left\|\nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta})\right\|_2 \leq \tau \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_2, \tag{2.6.1}$$

for any $\boldsymbol{\Theta} \in \mathbb{B}_\alpha(\boldsymbol{\Theta}^*)$. Remind that $\boldsymbol{\Theta}_k^{'} = \vec{(}\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) \in \mathbb{R}^{p^2+p}$. According to the derivation in the proof of Lemma 2.2.1, we have

$$\nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\Theta}_k^{'}|\boldsymbol{\Theta}) = \begin{pmatrix} \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\Omega}_k^{'}(\boldsymbol{X}-\boldsymbol{\mu}_k^{'})\right] \\ \text{vec}\left\{\frac{1}{2}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\boldsymbol{\Omega}_k^{'-1} - \frac{1}{2}\mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X}-\boldsymbol{\mu}_k^{'})(\boldsymbol{X}-\boldsymbol{\mu}_k^{'})^\top]\right\}^\top \end{pmatrix}.$$

Define $D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta}) = L_{\boldsymbol{\Theta}^*,k}(\boldsymbol{X}) - L_{\boldsymbol{\Theta},k}(\boldsymbol{X})$. Therefore, the square of the left hand side of (2.6.1) can be simplified to

$$\begin{aligned} & \left\|\nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta})\right\|_2^2 \\ = & \underbrace{\left\|\mathbb{E}\left[D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})\boldsymbol{\Omega}_k^*(\boldsymbol{X}-\boldsymbol{\mu}_k^*)\right]\right\|_2^2}_{I} \\ & + \underbrace{\left\|\frac{1}{2}\mathbb{E}[D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})\boldsymbol{\Omega}_k^{*-1} - \frac{1}{2}\mathbb{E}[D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})(\boldsymbol{X}-\boldsymbol{\mu}_k^*)(\boldsymbol{X}-\boldsymbol{\mu}_k^*)^\top]\right\|_F^2}_{II}. \end{aligned}$$

If we can show $I \leq \tau_1\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_2^2$ and $II \leq \tau_2\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_2^2$, then we have $\tau = \sqrt{\tau_1+\tau_2}$ since

$$\left\|\nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\Theta}_k^{'}} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*|\boldsymbol{\Theta})\right\|_2 \leq \sqrt{\tau_1+\tau_2}\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_2.$$

Bounding I: We apply Taylor expansion to simplify $D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})$. Remind that, by assumption, $\pi_k = 1/K$, and hence we have

$$L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) = \frac{\pi_k f_k(\boldsymbol{X};\boldsymbol{\Theta}_k)}{\sum_{k=1}^K \pi_k f_k(\boldsymbol{X};\boldsymbol{\Theta}_k)} = \frac{|\boldsymbol{\Omega}_k|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k(\boldsymbol{X}-\boldsymbol{\mu}_k)\right\}}{\sum_{k=1}^K |\boldsymbol{\Omega}_k|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k(\boldsymbol{X}-\boldsymbol{\mu}_k)\right\}}.$$

Then, Taylor expansion of $L_{\boldsymbol{\Theta},k}(\boldsymbol{X})$ around $\boldsymbol{\Theta}_k^*$ leads to

$$L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) = L_{\boldsymbol{\Theta}^*,k}(\boldsymbol{X}) + [\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})]^\top(\boldsymbol{\Theta}-\boldsymbol{\Theta}^*), \tag{2.6.2}$$

where $\boldsymbol{\Theta}_t = \boldsymbol{\Theta}^* + t\Delta$ with $t \in [0,1]$ and $\Delta = \boldsymbol{\Theta}-\boldsymbol{\Theta}^*$. Here the derivative of $L_{\boldsymbol{\Theta},k}(\boldsymbol{X})$ with respect to $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K)$ can be written as

$$\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) = \left([\nabla_{\boldsymbol{\Theta}_1} L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]^\top, \ldots, [\nabla_{\boldsymbol{\Theta}_K} L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]^\top\right)^\top, \tag{2.6.3}$$

where

$$\nabla_{\boldsymbol{\Theta}_j} L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) = \begin{cases} -L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \cdot L_{\boldsymbol{\Theta},j}(\boldsymbol{X}) \cdot \delta_{\boldsymbol{\Theta}_j}(\boldsymbol{X}) \text{ when } j \neq k; \\ L_{\boldsymbol{\Theta},k}(\boldsymbol{X})[1-L_{\boldsymbol{\Theta},k}(\boldsymbol{X})] \cdot \delta_{\boldsymbol{\Theta}_k}(\boldsymbol{X}) \text{ when } j = k, \end{cases}$$

and, for $j = 1 \ldots, K$, and $\boldsymbol{\Theta}_j = \vec{(}\mu_j, \boldsymbol{\Omega}_j)$,

$$\delta_{\boldsymbol{\Theta}_j}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{\Omega}_j(\boldsymbol{X} - \boldsymbol{\mu}_j) \\ \frac{1}{2}\mathrm{vec}\left\{\boldsymbol{\Omega}_j^{-1} - (\boldsymbol{X} - \boldsymbol{\mu}_j)(\boldsymbol{X} - \boldsymbol{\mu}_j)^\top\right\} \end{pmatrix}.$$

Next we apply this Taylor expansion to bound I. According to (2.6.2), we have

$$\begin{aligned} I &= \left\|\mathbb{E}\left[\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)[\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})]^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\right]\right\|_2^2 \\ &= \left\|\mathbb{E}\left[\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)[\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})]^\top\right]\right\|_2^2 \cdot \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2^2 \\ &\leq \underbrace{\sup_{t\in[0,1]} \mathbb{E}\left[\|\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\|_2^2 \cdot \|\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\|_2^2\right]}_{\tau_1} \cdot \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2^2. \end{aligned}$$

By the definition of $\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})$, which equals to (2.6.3) with $\boldsymbol{\Theta} = \boldsymbol{\Theta}_t$, we have

$$\begin{aligned} \|\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\|_2^2 &= \underbrace{\sum_{j\neq k} [L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X}) L_{\boldsymbol{\Theta}_t,j}(\boldsymbol{X})]^2 \cdot [\delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X})]^\top \delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X})}_{A_1} \\ &\quad + \underbrace{\left[L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\big(1 - L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\big)\right]^2 \cdot [\delta_{\boldsymbol{\Theta}_{tk}}(\boldsymbol{X})]^\top \delta_{\boldsymbol{\Theta}_{tk}}(\boldsymbol{X})}_{A_2}. \end{aligned}$$

For each $j = 1, \ldots, K$, we define

$$W_j := \sup_{t\in[0,1]} \mathbb{E}\left\{[\delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X})]^\top \delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X}) \cdot \|\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\|_2^2\right\}, \tag{2.6.4}$$

Then

$$\tau_1 \leq \sup_{t\in[0,1]} \mathbb{E}\left[\|\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\|_2^2 (A_1 + A_2)\right]. \tag{2.6.5}$$

Under Condition 2.2.2, it is sufficient to get an upper bound for τ_1,

$$\begin{aligned} \tau_1 &\leq \sup_{t\in[0,1]} \mathbb{E}\left[\|\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\|_2^2 A_1\right] + \sup_{t\in[0,1]} \mathbb{E}\left[\|\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\|_2^2 A_2\right] \\ &\leq \sum_{j\neq k} \frac{\gamma^2}{24^2(K-1)^2 M_j} \cdot W_j + \left(\frac{\gamma}{24(K-1)\sqrt{M_k}}(K-1)\right)^2 \cdot W_k. \end{aligned}$$

It implies that

$$\tau_1 \leq \frac{\gamma^2}{288}. \tag{2.6.6}$$

Bounding II: We can apply similar trick above to bound II. By triangle inequality, we have

$$\begin{aligned} II &\leq \underbrace{\left\| \frac{1}{2}\mathbb{E}\big[D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})\boldsymbol{\Omega}_k^{*-1}\big] \right\|_F^2}_{II_1} \\ &\quad + \underbrace{\left\| \frac{1}{2}\mathbb{E}\big[D_L(\boldsymbol{\Theta}^*, \boldsymbol{\Theta})(\boldsymbol{X} - \boldsymbol{\mu}_k^*)(\boldsymbol{X} - \boldsymbol{\mu}_k^*)^\top\big] \right\|_F^2}_{II_2}. \end{aligned}$$

Apply Taylor expansion in (2.6.2), we obtain

$$\begin{aligned} II_1 &\leq \underbrace{\frac{1}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\|_2^2 \, \|\boldsymbol{\Omega}_k^{*-1}\|_F^2\right]}_{\gamma_{21}} \cdot \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2^2 \\ II_2 &\leq \underbrace{\frac{1}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{\Theta}} L_{\boldsymbol{\Theta}_t,k}(\boldsymbol{X})\|_2^2 \, \left\|(\boldsymbol{X} - \boldsymbol{\mu}_k^*)(\boldsymbol{X} - \boldsymbol{\mu}_k^*)^\top\right\|_F^2\right]}_{\gamma_{22}} \cdot \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2^2. \end{aligned}$$

Analogously to (2.6.4), we define

$$W_j' := \sup_{t\in[0,1]} \mathbb{E}\left\{ [\delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X})]^\top \delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X}) \left\|\boldsymbol{\Omega}_k^{*-1}\right\|_F^2 \right\}, \tag{2.6.7}$$

$$W_j'' := \sup_{t\in[0,1]} \mathbb{E}\left\{ [\delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X})]^\top \delta_{\boldsymbol{\Theta}_{tj}}(\boldsymbol{X}) \left\|(\boldsymbol{X} - \boldsymbol{\mu}_k^*)(\boldsymbol{X} - \boldsymbol{\mu}_k^*)^\top\right\|_F^2 \right\}. \tag{2.6.8}$$

for each $j = 1, \ldots, K$. Under Condition 2.2.2, we have that,

$$\tau_{21} < \frac{\gamma^2}{576}, \quad \tau_{22} < \frac{\gamma^2}{576}, \text{ and hence } \tau_2 < \frac{\gamma^2}{288}.$$

This together with (2.6.6) implies that $\tau = \sqrt{\tau_1 + \tau_2} < \gamma/12$, namely

$$\left\| \nabla_{\boldsymbol{\Theta}_k'} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^* | \boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\Theta}_k'} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^* | \boldsymbol{\Theta}) \right\|_2 \leq \frac{\gamma}{12}.$$

Now we take the summation

$$\sum_{k=1}^K \left\| \nabla_{\boldsymbol{\Theta}_k'} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^* | \boldsymbol{\Theta}^*) - \nabla_{\boldsymbol{\Theta}_k'} Q(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^* | \boldsymbol{\Theta}) \right\|_2^2 \leq \frac{\gamma}{12} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2, \tag{2.6.9}$$

for any $\boldsymbol{\Theta} \in \mathbb{B}_\alpha(\boldsymbol{\Theta}^*)$. This ends the proof of Lemma 2.2.3. ■

2.6.3 Proof of Lemma 2.2.5

In order to compute γ, we consider each $\mathbf{\Theta}_k = \{\mu_k, \mathbf{\Omega}_k\}$ individually. That means we prove the following part first:

$$Q_n(\mathbf{\Theta}_k'|\mathbf{\Theta}) - Q_n(\mathbf{\Theta}_k^*|\mathbf{\Theta}) - \left\langle \nabla Q_n(\mathbf{\Theta}_k^*|\mathbf{\Theta}), \mathbf{\Theta}_k' - \mathbf{\Theta}_k^* \right\rangle \leq -\frac{\gamma}{2} \left\| \mathbf{\Theta}_k' - \mathbf{\Theta}_k^* \right\|_2^2,$$

where $Q_n(\mathbf{\Theta}_k|\mathbf{\Theta})$ means we set $\mathbf{\Theta}_i$ $i \neq k$ to zero.

It is sufficient to compute γ_k in (2.2.9). Remind that $\mathbf{\Theta}_k^{'} = \vec{(}\boldsymbol{\mu}_k, \mathbf{\Omega}_k) \in \mathbb{R}^{p^2+p}$. Therefore,

$$\nabla_{\mathbf{\Theta}_k^{'}} Q_n(\mathbf{\Theta}_k^{'}|\mathbf{\Theta}) = ([\nabla_{\boldsymbol{\mu}_k^{'}} Q_n(\mathbf{\Theta}_k^{'}|\mathbf{\Theta})]^\top, [\mathrm{vec}(\nabla_{\mathbf{\Omega}_k^{'}} Q_n(\mathbf{\Theta}_k^{'}|\mathbf{\Theta}))]^\top)^\top, \tag{2.6.10}$$

with

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}_k^{'}} Q_n(\mathbf{\Theta}_k^{'}|\mathbf{\Theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)\mathbf{\Omega}_k^{'}(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{'})\right] \\
\nabla_{\mathbf{\Omega}_k^{'}} Q_n(\mathbf{\Theta}_k^{'}|\mathbf{\Theta}) &= \frac{1}{2n}\sum_{i=1}^{n}[L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)]\mathbf{\Omega}_k^{'-1} \\
&- \frac{1}{2n}\sum_{i=1}^{n}[L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{'})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{'})^\top].
\end{aligned}$$

Denote $h(\boldsymbol{\mu}, \mathbf{\Omega}) := \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \mathbf{\Omega}(\boldsymbol{x}_i - \boldsymbol{\mu})$. According to the definition in (2.1.8), we have

$$\begin{aligned}
Q_n(\mathbf{\Theta}_k'|\mathbf{\Theta}) - Q_n(\mathbf{\Theta}_k^*|\mathbf{\Theta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)\big\{\frac{1}{2}\log\det(\mathbf{\Omega}_k') \\
&\quad -\frac{1}{2}\log\det(\mathbf{\Omega}_k^*) + h(\boldsymbol{\mu}_k^*, \mathbf{\Omega}_k^*) - h(\boldsymbol{\mu}_k', \mathbf{\Omega}_k')\big\}\Big].
\end{aligned}$$

This together with (2.6.10) implies that

$$Q_n(\mathbf{\Theta}_k'|\mathbf{\Theta}) - Q_n(\mathbf{\Theta}_k^*|\mathbf{\Theta}) - \left\langle \nabla_{\mathbf{\Theta}_k^{'}} Q_n(\mathbf{\Theta}_k^*|\mathbf{\Theta}), \mathbf{\Theta}_k' - \mathbf{\Theta}_k^* \right\rangle = I + II,$$

where

$$I = \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\left\{h(\boldsymbol{\mu}_k^*,\boldsymbol{\Omega}_k^*) - h(\boldsymbol{\mu}_k',\boldsymbol{\Omega}_k^*)\right\}\right]$$
$$-(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*)^\top \nabla_{\boldsymbol{\mu}_k'} Q_n(\boldsymbol{\Theta}_k^*|\boldsymbol{\Theta}^{(t)}),$$
$$II = \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\left\{\frac{1}{2}\log\det(\boldsymbol{\Omega}_k') - \frac{1}{2}\log\det(\boldsymbol{\Omega}_k^*)\right.\right.$$
$$\left.\left.+h(\boldsymbol{\mu}_k',\boldsymbol{\Omega}_k^*) - h(\boldsymbol{\mu}_k',\boldsymbol{\Omega}_k')\right\}\right] - [\mathrm{vec}(\boldsymbol{\Omega}_k' - \boldsymbol{\Omega}_k^*)]^\top \nabla_{\boldsymbol{\Omega}_k'} Q_n(\boldsymbol{\Theta}_k^*|\boldsymbol{\Theta}^{(t)}).$$

By a little algebra, we can show that

$$I = -\frac{1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Omega}_k^*(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*).$$

Due to the positive definiteness of $\boldsymbol{\Omega}_k^*$, it is shown the following inequality

$$(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*)^\top(\boldsymbol{\Omega}_k^* - \sigma_{\min}(\boldsymbol{\Omega}_k^*)I_p)(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*) \geq 0$$

$$(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Omega}_k^*(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*) \geq (\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*)^\top \sigma_{\min}(\boldsymbol{\Omega}_k^*) I_p(\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*) \geq \beta_1 \left\|\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*\right\|_2^2.$$

Substituting the above bound, it is shown that

$$I \leq -\frac{\beta_1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\|\boldsymbol{\mu}_k' - \boldsymbol{\mu}_k^*\|_2^2. \tag{2.6.11}$$

Therefore, it remains to show that

$$II \leq -\frac{1}{2n}\sum_{i=1}^{n}\frac{L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)}{2(\beta_2+2\alpha)^2}\|\mathrm{vec}(\boldsymbol{\Omega}_k' - \boldsymbol{\Omega}_k^*)\|_2^2. \tag{2.6.12}$$

Note that, in order to show (2.6.12), it is equivalent to deriving the strong concavity parameter of $g(\boldsymbol{\Omega}_k)$, where

$$g(\boldsymbol{\Omega}_k) := \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\left\{\frac{1}{2}\log\det(\boldsymbol{\Omega}_k) - h(\boldsymbol{\mu}_k',\boldsymbol{\Omega}_k)\right\}\right].$$

To see it, finding the strong concavity parameter of $g(\boldsymbol{\Omega}_k)$ aims to compute ρ_k such that, for any $\boldsymbol{\Omega}_k', \boldsymbol{\Omega}_k^* \in \mathcal{B}_\alpha(\boldsymbol{\Omega}_k^*)$,

$$g(\boldsymbol{\Omega}_k') - g(\boldsymbol{\Omega}_k^*) - \left\langle \mathrm{vec}\left(\nabla g(\boldsymbol{\Omega}_k^*)\right), \mathrm{vec}(\boldsymbol{\Omega}_k' - \boldsymbol{\Omega}_k^*)\right\rangle \leq -\rho_k/2 \cdot \|\boldsymbol{\Omega}_k' - \boldsymbol{\Omega}_k^*\|_F^2,$$

where the left hand side is exactly II. According to Taylor expansion, we can expand $g(\mathbf{\Omega}_k')$ around $\mathbf{\Omega}_k^*$ and obtain

$$\begin{aligned} g(\mathbf{\Omega}_k') &= g(\mathbf{\Omega}_k^*) + \left\langle \mathrm{vec}(\nabla g(\mathbf{\Omega}_k^*), \mathrm{vec}(\mathbf{\Omega}_k' - \mathbf{\Omega}_k^*) \right\rangle \\ &\quad + \frac{1}{2} \left[\mathrm{vec}(\mathbf{\Omega}_k' - \mathbf{\Omega}_k^*)\right]^\top \nabla^2 g(\mathbf{Z}) \left[\mathrm{vec}(\mathbf{\Omega}_k' - \mathbf{\Omega}_k^*)\right], \end{aligned}$$

where $\mathbf{Z} = t\mathbf{\Omega}_k' + (1-t)\mathbf{\Omega}_k^*$ with $t \in [0,1]$. For any two matrices $\mathbf{A}, \mathbf{B}$, we write $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. We denote $\mathbb{1}_p$ as the identity matrix with dimension $p \times p$. And $\sigma_i(A)$ is the i-th eigenvalue of matrix A. Therefore, if we can show that $-\nabla^2 g(\mathbf{Z}) \succeq m\, \mathbb{1}_p$, i.e., the minimal eigenvalue value $\sigma_{\min}(-\nabla^2 g(\mathbf{Z})) \geq m$, for some positive $m \in \mathbb{R}$, then we have the strongly concavity parameter $\rho_k = m$. By the definition, we have $\nabla^2 g(\mathbf{\Omega}_k^*) = -\frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)[\mathbf{\Omega}_k^*]^{-1} \otimes [\mathbf{\Omega}_k^*]^{-1}$. Denote $\widetilde{\Delta} = \mathbf{\Omega}_k' - \mathbf{\Omega}_k^*$. We obtain

$$-\nabla^2 g(\mathbf{Z}) = \frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i) \left(\mathbf{\Omega}_k^* + t\widetilde{\Delta}\right)^{-1} \otimes \left(\mathbf{\Omega}_k^* + t\widetilde{\Delta}\right)^{-1}.$$

According to Theorem 4.2.1 2 in (81), for any two matrices $\mathbf{A}, \mathbf{B}$, the minimal eigenvalue value of $\mathbf{A} \otimes \mathbf{B}$ equals the products of the minimal eigenvalue values of $\mathbf{A}$ and $\mathbf{B}$. Therefore, we have $\sigma_{\min}\left(\mathbf{A}^{-1} \otimes \mathbf{A}^{-1}\right) = \left[\sigma_{\min}(\mathbf{A}^{-1})\right]^2 = \left[\sigma_{\max}(\mathbf{A})\right]^{-2} = \|\mathbf{A}\|_2^{-2}$, where $\|\mathbf{A}\|_2$ refers to the spectral norm of matrix $\mathbf{A}$. Hence,

$$\begin{aligned} \sigma_{\min}(-\nabla^2 g(\mathbf{Z})) &= \frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i) \|\mathbf{\Omega}_k^* + t\widetilde{\Delta}\|_2^{-2} \\ &\geq \frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i) \left[\|\mathbf{\Omega}_k^*\|_2 + \|t\widetilde{\Delta}\|_2\right]^{-2}. \end{aligned}$$

As $\|\mathbf{\Theta}' - \mathbf{\Theta}^*\| \leq 2\alpha$, $\|\mathbf{\Omega}_k' - \mathbf{\Omega}_k^*\|_2 \leq \|\mathbf{\Theta}' - \mathbf{\Theta}^*\|_2 \leq 2\alpha$. Therefore,

$$\begin{aligned} \sigma_{\min}(-\nabla^2 g(\mathbf{Z})) &\geq \frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i) \left[\|\mathbf{\Omega}_k^*\|_2 + 2\alpha\right]^{-2} \\ &\geq \frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i) \left(\beta_2 + 2\alpha\right)^{-2}, \end{aligned}$$

which implies (2.6.12). Putting the upper bound of I and II together,

$$I + II \leq -\underbrace{\frac{1}{2n}\sum_{i=1}^n L_{\mathbf{\Theta},k}(\boldsymbol{x}_i)}_{(a)} \cdot \min\left\{\beta_1, \frac{1}{2(\beta_2 + 2\alpha)^2}\right\} \|\mathbf{\Theta}_k' - \mathbf{\Theta}_k^*\|_2^2. \quad (2.6.13)$$

However, (a) is a random term but we require a non-random strong concavity parameter. Thus a concentration bound will be applied on it. $\{L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i), i = 1, \ldots, n\}$ are independent random variables with $0 \leq L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) \leq 1$. After applying a basic Hoeffding's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\right| \leq t\right) \geq 1 - 2e^{-2nt^2},$$

which implies

$$\left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\right| \leq \sqrt{\frac{1}{2}\log\frac{2K}{\delta}}\sqrt{\frac{1}{n}},$$

with probability at least $1 - \delta/K$. As $\sqrt{\log(2K/\delta)/2n} = o(1)$, there exists some constant c such that

$$\sqrt{\frac{\log 2K}{2\delta n}} - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})] \leq -c,$$

when n is large enough. Then plugging it into (2.6.13),

$$I + II \leq -\frac{1}{2}c \cdot \min\left\{\beta_1, \frac{1}{2(\beta_2 + 2\alpha)^2}\right\} \|\boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k\|_2^2,$$

with probability at least $1 - \delta/K$, where

$$\gamma = c\min\left\{\beta_1, \frac{1}{2(\beta_2 + 2\alpha)^2}\right\}.$$

Once the individual strong concavity parameter is computed, we can simply take the summation from 1 to K:

$$\sum_{k=1}^{K} Q_n(\boldsymbol{\Theta}'_k|\boldsymbol{\Theta}) - Q_n(\boldsymbol{\Theta}^*_k|\boldsymbol{\Theta}) - \langle\nabla Q_n(\boldsymbol{\Theta}^*_k|\boldsymbol{\Theta}), \boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k\rangle \leq -\frac{1}{2}\sum_{k=1}^{K}\gamma\|\boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k\|_2^2$$

which implies

$$Q_n(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) - Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \left\langle\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}), \boldsymbol{\Theta}' - \boldsymbol{\Theta}^*\right\rangle \leq -\frac{1}{2}\gamma\|\boldsymbol{\Theta}' - \boldsymbol{\Theta}^*\|_2^2$$

with probability at least $1 - \delta$. This ends the proof of Lemma 2.2.5. ■

2.6.4 A Key Lemma for Proving Corollary 2.2.12

The next lemma computes the statistical errors in Condition 2.2.6 for our SCAN penalty and provides explicit forms of the corresponding $\varepsilon_1, \varepsilon_2$ and δ_1, δ_2.

Lemma 2.6.1. Suppose that Condition 2.2.10, 2.2.11 hold, then Condition 2.2.6 is satisfied for SCAN penalty with

$$\varepsilon_1 = (CK\|\boldsymbol{\Omega}^*\|_\infty + C'K^{1.5})\sqrt{\frac{\log p + \log(e/\delta)}{n}}, \delta_1 = (18K^2 + 6K)\delta, \tag{2.6.14}$$

$$\varepsilon_2 = C''\sqrt{p}\sqrt{\frac{K^3\left(\log p + \log(e/\delta)\right)}{n}}, \delta_2 = (8K^2 + 2K)\delta, \tag{2.6.15}$$

for some absolute constant $C, C', C'' > 0$. Here $\|\boldsymbol{\Omega}^*\|_\infty$ is the overall max induced norm defined as $\|\boldsymbol{\Omega}^*\|_\infty = \max_{k\in[K]} \|\boldsymbol{\Omega}_k^*\|_\infty$.

In Lemma 2.6.1, the number of clusters K is allowed to grow with the sample size n and the dimension p. The diverging rate of K controls the convergence probability at each iteration and is upper bounded to ensure that the statistical errors hold with a high probability tending to 1 with a proper choice of δ, e.g., $\delta = 1/p$.

Proof of Lemma 2.6.1: For the first part of this proof, we focus on the upper bound of $\|\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\|_{\mathcal{P}^*}$. Recall that

$$\begin{aligned}
&\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) \\
=& \begin{pmatrix} \nabla_{\boldsymbol{\Theta}_1^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Theta}_1^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) \\ \vdots \\ \nabla_{\boldsymbol{\Theta}_K^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Theta}_K^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) \end{pmatrix} \\
=& \begin{pmatrix} \nabla_{\boldsymbol{\mu}_1^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\mu}_1^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) \\ \mathrm{vec}\left\{\nabla_{\boldsymbol{\Omega}_1^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_1^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right\}^\top \\ \vdots \\ \nabla_{\boldsymbol{\mu}_K^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\mu}_K^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) \\ \mathrm{vec}\left\{\nabla_{\boldsymbol{\Omega}_K^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_K^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right\}^\top \end{pmatrix}.
\end{aligned} \tag{2.6.16}$$

For simplicity, we define $h_{\mu_k}(\boldsymbol{\Theta}^*) = \nabla_{\boldsymbol{\mu}_k^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\mu}_k^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})$ and $h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*) = \nabla_{\boldsymbol{\Omega}_k^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_k^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})$. Then from the definition of dual norm $\mathcal{P}^*$ (2.5.1), we can have

$$\|\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\|_{\mathcal{P}^*} \leq M_1 \max_{k\in[K]} \underbrace{\|h_{\mu_k}(\boldsymbol{\Theta}^*)\|_\infty}_{I}$$

$$+M_2 \max_{k\in[K]} \underbrace{\left\|h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*)\right\|_{\max}}_{II} + M_3 \underbrace{\max_{i,j} \left\| \left[h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*)\right]_{ij}, \ldots, \left[h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*)\right]_{ij} \right\|_2}_{III},$$

which are corresponding to the penalty on element-wise cluster means, element-wise precision matrices and group structures of multiple precision matrices, respectively.

Bounding Statistical Error for k-th Cluster Mean: Referring to the proof in Lemma 2.2.1,

$$h_{\mu_k^*}(\boldsymbol{\Theta}^*) \;=\; \frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{\Omega}_k^*(\boldsymbol{x}_i - \boldsymbol{\mu}_k^*) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\Omega}_k^*(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\right].$$

Note that $\|\boldsymbol{\Omega}_k^*\|_\infty$ is a scalar. By using triangle inequality, we simplify I by two parts:

$$\begin{aligned} I \;&\leq\; \|\boldsymbol{\Omega}_k^*\|_\infty \left\| \frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^*) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{\mu}_k^*)\right] \right\|_\infty \\ &\leq\; \|\boldsymbol{\Omega}_k^*\|_\infty \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}\right] \right\|_\infty}_{I_1} \\ &\quad + \|\boldsymbol{\Omega}_k^*\|_\infty \underbrace{\left\| \left(\frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\right] \right) \boldsymbol{\mu}_k^* \right\|_\infty}_{I_2}. \end{aligned}$$

Bounding I_1: Denote

$$\zeta = \frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}\right]$$

For $\zeta \in \mathbb{R}^p$, we consider the j-th coordinate ζ_j of ζ

$$\zeta_j = \frac{1}{n}\sum_{i=1}^n L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)x_{ij} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})X_j\right]. \tag{2.6.17}$$

We introduce a set of missing data $\{c_i, i = 1, \ldots, n\}$, which are independent copies of random variable c. The pair $(\boldsymbol{x}_i, c_i)$ are the independent copy of $(\boldsymbol{X}, c)$. Here c

takes a value from the set $\{1, \ldots, K\}$, where $c = k'$ indicates that $\boldsymbol{X}$ was generated by the k'-th mixture component. In another word, the conditional distribution of $\boldsymbol{X}$ is defined below:

$$\boldsymbol{X}|c = k' \sim \mathcal{N}(\boldsymbol{\mu}^*_{k'}, \boldsymbol{\Sigma}^*_{k'})$$
$$\mathbb{P}(c = k') = \pi_{k'}, \ \sum_{k'}^{K} \pi_{k'} = 1.$$

This is the usual choice of missing data in EM approaches to mixture modeling. The quantity $(\boldsymbol{x}_i, c_i)$ is referred to as the completed data. Now by the assumption, the j-th coordinate x_{ij} of $\boldsymbol{x}_i$ can be rewritten as the form below:

$$x_{ij} = \sum_{k'=1}^{K} I\{c_i = k'\}(\mu^*_{k'j} + V_{k'j}), \ j \in [p] \tag{2.6.18}$$

where $\mu^*_{k'j}$ is the j-th coordinate of the true cluster mean $\boldsymbol{\mu}^*_{k'}$ and $V_{k'j} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^*_{k'jj})$. Plugging (2.6.18) into (2.6.17), it suffices to bound ζ_j.

$$\begin{aligned}
|\zeta_j| \ &\leq \ \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k'=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j} - \mathbb{E}\left[\sum_{k'=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} \mu^*_{k'j} \right] \right| \\
&+ \ \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k'=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} V^*_{k'j} - \mathbb{E}\left[\sum_{k'=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} V^*_{k'j} \right] \right| \\
&\leq \ \sum_{k'=1}^{K} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j} - \mathbb{E}\Big[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} \mu^*_{k'j} \Big] \right|}_{\zeta_{j_1}} \\
&+ \ \sum_{k'=1}^{K} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} V^*_{k'j} - \mathbb{E}\Big[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} V^*_{k'j} \Big] \right|}_{\zeta_{j_2}}.
\end{aligned}$$

We bound ζ_{j_1} first. Based on the fact that $|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j}| \leq |\mu^*_{k'j}| \leq \|\boldsymbol{\mu}^*_{k'}\|_\infty$ almost surely it can show that $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j}$ is a sub-gaussian random variable with norm $\|\boldsymbol{\mu}^*_{k'}\|_\infty$. Following the Example 5.8 in (82), $\|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j}\|_{\psi_2} \leq \|\boldsymbol{\mu}^*_{k'}\|_\infty$ where $\|\cdot\|_{\psi_2}$ is defined as sub-Gaussian norm. According to supporting Lemma 2.8.3

$$\left\| L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} \mu^*_{k'j} \right] \right\|_{\psi_2} \leq 2 \left\| \boldsymbol{\mu}^*_{k'} \right\|_\infty.$$

The standard concentration result in supporting Lemma 2.8.4 yields that for every $t \geq 0$ and some constant D_1,

$$\mathbb{P}\left(|\zeta_{j_1}| \geq t\right) \leq e \exp\left(-\frac{D_1 n t^2}{4\|\boldsymbol{\mu}^*_{k'}\|^2_\infty}\right),$$

which implies that, with probability at least $1-\delta$,

$$|\zeta_{j_1}| \leq \sqrt{\frac{4}{D_1}}\|\boldsymbol{\mu}^*_{k'}\|_\infty \sqrt{\frac{\log(e/\delta)}{n}}. \tag{2.6.19}$$

Now we start to bound ζ_{j_2}. The fact that $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\} \leq 1$ shows that it is a sub-gaussian random variable with norm $\|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}\|_{\psi_2} \leq 1$. $V^*_{k'j}$ is a Gaussian random variable so that it is also a sub-gaussian random variable with norm $\|V^*_{k'j}\|_{\psi_2} \leq (\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2}$. Then using the result in supporting Lemma 2.8.2, $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}V^*_{k'j}$ is sub-exponential random variable. Moreover, there exists constant D_2 such that

$$\left\|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}V^*_{k'j}\right\|_{\psi_1} \leq D_2\Big(\|\boldsymbol{\Sigma}^*_{k'}\|_{\max}\Big)^{1/2}.$$

Supporting lemma 2.8.3 implies

$$\left\|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}V^*_{k'j} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})I\{c = k'\}V^*_{k'j}\right]\right\|_{\psi_1} \leq 2D_2\Big(\|\boldsymbol{\Sigma}^*_{k'}\|_{\max}\Big)^{1/2}.$$

Following the concentration inequality of sub-exponential random variables in supporting Lemma 2.8.5, there exists some constant D_3 such that the following inequality

$$\mathbb{P}\Big(|\zeta_{j_2}| \geq t\Big) \leq 2\exp\left(-D_3 \min\left\{\frac{t^2}{4D_2^2\|\boldsymbol{\Sigma}^*_{k'}\|_{\max}}, \frac{t}{2D_2(\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2}}\right\} n\right),$$

holds every $t \geq 0$. For sufficient small t, it reduces to

$$\mathbb{P}\Big(|\zeta_{j_2}| \geq t\Big) \leq 2\exp\left(-D_3\frac{nt^2}{4D_2\|\boldsymbol{\Sigma}^*_{k'}\|_{\max}}\right),$$

which implies that

$$|\zeta_{j_2}| \leq \sqrt{\frac{4D_2}{D_3}}(\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2}\sqrt{\frac{\log(2/\delta)}{n}}, \tag{2.6.20}$$

with probability at least $1-\delta$.

Adding (2.6.19) and (2.6.20) together, we have

$$\begin{aligned}|\zeta_{j_1}| + |\zeta_{j_2}| &\leq \sqrt{\frac{4}{D_1}}\|\boldsymbol{\mu}_{k'}^*\|_\infty\sqrt{\frac{\log(e/\delta)}{n}} + \sqrt{\frac{4D_2}{D_3}}(\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\sqrt{\frac{\log(2/\delta)}{n}} \\ &\leq \sqrt{\frac{4}{D}}\left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\right)\sqrt{\frac{\log(e/\delta)}{n}},\end{aligned}$$

by taking $D = \min\{D_1, D_3/D_2\}$, with at least probability $1 - 2\delta$. Therefore, it's sufficient to bound $|\zeta_j|$ by

$$|\zeta_j| \leq \sqrt{\frac{4}{D}}\sum_{k'=1}^{K}\left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\right)\sqrt{\frac{\log(e/\delta)}{n}},$$

with at least probability $1 - 2K\delta$. Taking the union bound over p coordinates, we obtain

$$I_1 \leq \sqrt{\frac{4}{D}}\sum_{k'=1}^{K}\left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\right)\sqrt{\frac{\log(e/\delta) + \log p}{n}}, \tag{2.6.21}$$

with at least probability $1 - 2K\delta$.

Bounding I_2: Recall that

$$I_2 = \left\|\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\right]\right)\boldsymbol{\mu}_k^*\right\|_\infty \leq \left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\right]\right| \|\boldsymbol{\mu}_k^*\|_\infty.$$

$\{L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)|i = 1, \ldots n\}$ are bounded independent random variables within interval between 0 and 1. Then it follows Hoeffding's inequality in supporting Lemma 2.8.6 that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\right| \leq t\right) \geq 1 - 2e^{-2nt^2},$$

which implies

$$\left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})]\right| \leq \sqrt{\frac{1}{2}\log\frac{2}{\delta}}\cdot\sqrt{\frac{1}{n}}, \tag{2.6.22}$$

with probability at least $1 - \delta$. Combining with the reminder term $\|\boldsymbol{\mu}_k^*\|$,

$$I_2 \leq \sqrt{\frac{1}{2}\log\frac{2}{\delta}}\cdot\sqrt{\frac{1}{n}}\|\boldsymbol{\mu}_k^*\|_\infty. \tag{2.6.23}$$

Note that the bound in (2.6.21) is $O_P((\log p/n)^{1/2})$ while the bound in (2.6.23) is $O_P((1/n)^{1/2})$, there exists some constant D_4 such that $I_2 \leq D_4 I_1$. Consequently, we conclude that I is upper bounded by

$$I \leq (1+D_4)\|\boldsymbol{\Omega}_k^*\|_\infty \sqrt{\frac{4}{D}} \sum_{k'=1}^{K} \left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\right) \sqrt{\frac{\log(e/\delta)+\log p}{n}},$$

with probability at least $1-(2K+1)\delta$. For simplicity, let

$$\varphi_K = \sum_{k'=1}^{K} \left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2}\right), \; C_1 = \sqrt{\frac{4(1+D_4)^2}{D}}. \tag{2.6.24}$$

Applying union bound,

$$\max_{k\in[K]} I \leq C_1 \left\|\boldsymbol{\Omega}^*\right\|_\infty \varphi_K \sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.25}$$

with probability at least $1-K(2K+1)\delta$.

Bounding Statistical Error for k-th Precision Matrix: Referring to the proof in Lemma 2.2.1,

$$\begin{aligned} h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*) &= \frac{1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_k^* - \frac{1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)^\top \\ &- \frac{1}{2}\mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\right]\boldsymbol{\Sigma}_k^* + \frac{1}{2}\mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X}-\boldsymbol{\mu}_k^*)(\boldsymbol{X}-\boldsymbol{\mu}_k^*)^\top\right]. \end{aligned}$$

Now we get an explicit from for $h_{\boldsymbol{\Omega}_k^*}(\boldsymbol{\Theta}^*)$. Then II is decomposed as below:

$$\begin{aligned} II \leq & \underbrace{\left\|\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{\Sigma}_k^* - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\Sigma}_k^*\right]\right)\right\|_{\max}}_{II_1} \\ + & \underbrace{\left\|\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)^\top - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})(\boldsymbol{X}-\boldsymbol{\mu}_k^*)(\boldsymbol{X}-\boldsymbol{\mu}_k^*)^\top\right]\right)\right\|_{\max}}_{II_2}. \end{aligned}$$

The first term is easy to deal with: since $\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\right]$ is scalar by the definition of $L_{\boldsymbol{\Theta},k}(\boldsymbol{X})$ we can pull it out of the norm. Combining with the result in (2.6.22), the first term is upper bounded by

$$II_1 \leq \|\boldsymbol{\Sigma}_k^*\|_{\max} \sqrt{\frac{1}{2}\log\frac{2}{\delta}} \cdot \sqrt{\frac{1}{n}}, \tag{2.6.26}$$

with probability at least $1-\delta$.

For the second term II_2, it can be decomposed as four following terms:

$$\begin{aligned} II_2 \;\leq\; & \underbrace{\left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top \right] \right) \right\|_{\max}}_{II_{21}} \\ & + \underbrace{\left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{\mu}_k^{*T} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X}\boldsymbol{\mu}_k^{*\top} \right] \right) \right\|_{\max}}_{II_{22}} \\ & + \underbrace{\left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{\mu}_k^*\boldsymbol{x}_i^\top - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\mu}_k^*\boldsymbol{X}^\top \right] \right) \right\|_{\max}}_{II_{23}} \\ & + \underbrace{\left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{\mu}_k^*\boldsymbol{\mu}_k^{*\top} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{\mu}_k^*\boldsymbol{\mu}_k^{*\top} \right] \right) \right\|_{\max}}_{II_{24}}. \end{aligned}$$

For the bound of II_{22} and II_{23}, we can just simply pull the $\boldsymbol{\mu}_k^*$ out, which implies

$$\begin{aligned} II_{22} \;&=\; \left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X} \right] \right) \boldsymbol{\mu}_k^{*\top} \right\|_{\max} \\ &\leq\; \left\| \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\boldsymbol{x}_i - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})\boldsymbol{X} \right] \right) \right\|_{\infty} \left\| \boldsymbol{\mu}_k^* \right\|_\infty \\ &\overset{(a)}{\leq}\; \sqrt{\frac{4}{D}} \left\| \boldsymbol{\mu}_k^* \right\|_\infty \varphi_K \sqrt{\frac{\log(e/\delta) + \log p}{n}}, \end{aligned} \tag{2.6.27}$$

with probability at least $1-2K\delta$, where (a) follows (2.6.21).

Next we turn to bound II_{21}. Expand $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ to matrix form for convenient use

$$\boldsymbol{x}_i\boldsymbol{x}_i^\top = \begin{pmatrix} x_{i1}x_{i1} & \dots & x_{i1}x_{ip} \\ \vdots & \ddots & \vdots \\ x_{ip}x_{i1} & \dots & x_{ip}x_{ip} \end{pmatrix}.$$

Since we require a matrix max norm here, it suffices to bound II_{21} individually, namely

$$\zeta_{jj'} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)x_{ij}x_{ij'} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})X_jX_{j'} \right] \right).$$

Recall in (2.6.18) the j-th coordinate of $\boldsymbol{x}_i$ could be expressed as

$$x_{ij} = \sum_{k'=1}^{K} I\{c_i = k'\}(\mu^*_{k'j} + V_{k'j}).$$

By straightforward algebra,

$$\begin{aligned} x_{ij}x_{ij'} &= \sum_{k'=1}^{K} I\{c_i = k'\}(\mu^*_{k'j} + V_{k'j}) \cdot \sum_{k'=1}^{K} I\{c_i = k'\}(\mu^*_{k'j'} + V_{k'j'}) \\ &\overset{(a)}{=} \sum_{k'=1}^{K} I\{c_i = k'\}^2(\mu^*_{k'j} + V_{k'j})(\mu^*_{k'j'} + V_{k'j'}) \\ &= \sum_{k'=1}^{K} I\{c_i = k'\}\left(\mu^*_{k'j}\mu^*_{k'j'} + \mu^*_{k'j}V_{k'j'} + V_{k'j}\mu^*_{k'j'} + V_{k'j}V_{k'j'}\right), \end{aligned}$$

where (a) follows the fact that $I\{c_i = k\}I\{c_i = k'\} = 0$ for any $k \neq k'$. Consequently, we divide $\zeta_{jj'}$ into four parts:

$$\zeta_{jj'} = \frac{1}{2}\sum_{k'=1}^{K}\left(\zeta_{jj'}(\mu^*_{k'j}\mu^*_{k'j'}) + \zeta_{jj'}(\mu^*_{k'j}V_{k'j'}) + \zeta_{jj'}(V_{k'j}\mu^*_{k'j'}) + \zeta_{jj'}(V_{k'j}V_{k'j'})\right),$$

where

$$\begin{aligned} \zeta_{jj'}(\mu^*_{k'j}\mu^*_{k'j'}) &= \frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}\mu^*_{k'j}\mu^*_{k'j'} \\ &\quad -\mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X})I\{c = k'\}\mu^*_{k'j}\mu^*_{k'j'}\right]. \end{aligned}$$

Taking the supreme over set $[p]$ in terms of p, p',

$$\begin{aligned} \sup_{j,j'\in[p]}|\zeta_{jj'}| \leq\ & \sum_{k'=1}^{K}\underbrace{\left(\sup_{j,j'\in[p]}|\zeta_{jj'}(\mu^*_{k'j}\mu^*_{k'j'})|\right)}_{(i)} + \sum_{k'=1}^{K}\underbrace{\left(\sup_{j,j'\in[p]}|\zeta_{jj'}(\mu^*_{k'j}V_{k'j'})|\right)}_{(ii)} \\ & + \sum_{k'=1}^{K}\underbrace{\left(\sup_{j,j'\in[p]}|\zeta_{jj'}(V_{k'j}\mu^*_{k'j'})|\right)}_{(iii)} + \sum_{k'=1}^{K}\underbrace{\left(\sup_{j,j'\in[p]}|\zeta_{jj'}(V_{k'j}V_{k'j'})|\right)}_{(iv)}. \end{aligned}$$

We will bound (i), (ii), (iii) and (iv) sequentially. $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}\mu^*_{k'j}\mu^*_{k'j'}$ is a sub-gaussian random variable with

$$\|L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)I\{c_i = k'\}\mu^*_{k'j}\mu^*_{k'j'}\|_{\psi_2} \leq \|\boldsymbol{\mu}^*_{k'}\|^2_{\infty}.$$

According to supporting Lemma 2.8.3,

$$\left\| L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j} \mu^*_{k'j'} - \mathbb{E}[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) I\{c = k'\} \mu^*_{k'j} \mu^*_{k'j'}] \right\|_{\psi_2} \leq 2 \|\boldsymbol{\mu}^*_{k'}\|^2_\infty.$$

Applying concentration inequality in supporting Lemma 2.8.4 yields that

$$\mathbb{P}\left(|\zeta_{jj'}(\mu^*_{k'j} \mu^*_{k'j'})| \leq t \right) \geq 1 - e \exp\left(-\frac{D_4 n t^2}{4 \|\boldsymbol{\mu}^*_{k'}\|^4_\infty} \right), \tag{2.6.28}$$

for any $t > 0$ and some constant D_4. After properly choosing t,

$$(i) \leq \sqrt{\frac{4}{D_4}} \|\boldsymbol{\mu}^*_{k'}\|^2_\infty \sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.29}$$

with probability at least $1 - \delta$. Note that both $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} \mu^*_{k'j} V_{k'j'}$ and $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} V_{k'j'} \mu^*_{k'j}$ are sub-exponential random variables with norm $\|\boldsymbol{\mu}^*_{k'}\|_\infty (\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2}$. Similar to the step in (2.6.20),

$$\left| \zeta_{jj'}(\mu^*_{k'j} V_{k'j'}) \right| \leq \sqrt{\frac{4}{D_5}} \left(\|\boldsymbol{\mu}^*_{k'}\|_\infty (\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2} \right) \sqrt{\frac{\log(2/\delta)}{n}},$$

with at least probability $1 - \delta$. Taking the union bound, it is shown that

$$(ii), (iii) \leq \sqrt{\frac{4}{D_5}} \left(\|\boldsymbol{\mu}^*_{k'}\|_\infty (\|\boldsymbol{\Sigma}^*_{k'}\|_{\max})^{1/2} \right) \sqrt{\frac{\log p + \log(2/\delta)}{n}}, \tag{2.6.30}$$

with probability at least $1 - \delta$ for sufficient large n.

Lastly, the fact that both $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} V_{k'j}$ and $V_{k'j'}$ are sub-gaussian random variables implies $L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) I\{c_i = k'\} V_{k'j} V_{k'j'}$ is sub-exponential random variable with parameter $\|\boldsymbol{\Sigma}^*_{k'}\|_{\max}$. Applying concentration result, there exists some constant D_6 such that the following inequality

$$\mathbb{P}\left(|\zeta_{jj'}(V_{k'j} V_{k'j'})| \geq t \right) \leq 2 \exp\left(-\frac{D_6 n t^2}{4 \|\boldsymbol{\Sigma}^*_{k'}\|^2_{\max}} \right),$$

holds for sufficiently small $t > 0$. Therefore,

$$\mathbb{P}\left(\sup_{j,j' \in [p]} |\zeta_{jj'}(V_{k'j} V_{k'j'})| \geq t \right) \leq 2p^2 \exp\left(-\frac{D_6 n t^2}{4 \|\boldsymbol{\Sigma}^*_{k'}\|^2_{\max}} \right).$$

When n is sufficiently large, with probability at least $1 - \delta$

$$(iv) \leq \sqrt{\frac{4}{D_6}} \|\boldsymbol{\Sigma}^*_{k'}\|_{\max} \sqrt{\frac{2 \log p + \log(2/\delta)}{n}}. \tag{2.6.31}$$

Putting (2.6.29), (2.6.30) and (2.6.31) together and after some adjustments, II_{21} is upper bounded by

$$II_{21} \leq \sqrt{\frac{1}{D_7}} \sum_{k'=1}^{K} \left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2} \right)^2 \sqrt{\frac{2\log p + \log(e/\delta)}{n}},$$

with probability at least $1 - 4K\delta$. $D_7 = \min(D_4, D_5, D_6)$. For simplicity, we denote

$$\varphi'_K = \sum_{k'=1}^{K} \left(\|\boldsymbol{\mu}_{k'}^*\|_\infty + (\|\boldsymbol{\Sigma}_{k'}^*\|_{\max})^{1/2} \right)^2.$$

Therefore,

$$II_{21} \leq \sqrt{\frac{2}{D_7}} \varphi'_K \sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.32}$$

with probability at least $1 - 4K\delta$.

For the last, it remains to bound II_{24}. Recall that

$$\begin{aligned} II_{24} &= \left\| \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) \boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*\top} - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*\top} \right] \right) \right\|_{\max} \\ &\leq \left| \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i) - \mathbb{E}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{X}) \right] \right) \right| \left\| \boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*\top} \right\|_{\max}. \end{aligned}$$

Applying the result in (2.6.22), we have

$$II_{24} \leq \|\boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*\top}\|_{\max} \sqrt{\frac{1}{2} \log \frac{2}{\delta}} \cdot \sqrt{\frac{1}{n}}, \tag{2.6.33}$$

with probability at least $1 - \delta$.

Putting (2.6.27), (2.6.32) and (2.6.33) together, now we can have a upper bound for II_2.

$$II_2 \leq \sqrt{\frac{1}{D_7}} \left(2\|\boldsymbol{\mu}_k^*\|_\infty \varphi_K + \varphi'_K \right) \sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.34}$$

for $D_7 < D/2$ with at least probability $1 - (8K+1)\delta$. The upper bound in (2.6.26) is of order $O_P(n^{-1/2})$ while the upper bound in (2.6.34) is of order $O_P((\log p/n)^{1/2})$. Thus there exists some constant D_8 such that $II_1 \leq D_8 II_2$. Let $C_2 = ((1 + D_8)^2/D_7)^{1/2}$. Applying union bound,

$$\max_{k \in [K]} II \leq C_2 \left(2\|\boldsymbol{\mu}^*\|_\infty \varphi_K + \varphi'_K \right) \sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.35}$$

with at least probability $1 - K(8K+2)\delta$.

Bound the Group Structure Part of Precision Matrix:

Recall that

$$\begin{aligned} III &= \max_{i,j} \Big\| \left[\nabla_{\boldsymbol{\Omega}_1^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_1^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right]_{ij}, \\ &\qquad\qquad \ldots, \left[\nabla_{\boldsymbol{\Omega}_K^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_K^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right]_{ij} \Big\|_2 \\ &\leq \max_{i,j} \sqrt{K} \Big\| \left[\nabla_{\boldsymbol{\Omega}_1^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_1^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right]_{ij}, \\ &\qquad\qquad \ldots, \left[\nabla_{\boldsymbol{\Omega}_K^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_K^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right]_{ij} \Big\|_\infty \\ &\leq \sqrt{K} \max_{k\in[K]} \left\| \left[\nabla_{\boldsymbol{\Omega}_k^*} Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Omega}_k^*} Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right] \right\|_{\max}. \end{aligned}$$

According to the result in (2.6.35) and applying union bound over $[K]$,

$$\mathbb{P}\left(III \geq C_2\sqrt{K}\left(2\|\boldsymbol{\mu}_k^*\|_\infty \varphi_K + \varphi'_K\right)\sqrt{\frac{\log p + \log(e/\delta)}{n}} \right) \leq K(8K+2)\delta.$$

Thus, III is upper bounded by

$$III \;\leq\; C_2\sqrt{K}\left(2\|\boldsymbol{\mu}^*\|_\infty \varphi_K + \varphi'_K\right)\sqrt{\frac{\log p + \log(e/\delta)}{n}}, \tag{2.6.36}$$

with at least probability $1 - K(8K+2)\delta$.

Finally, putting the upper bound (2.6.25), (2.6.35) and (2.6.36) together, we have a upper bound for the following statistical error

$$\begin{aligned} &\left\|\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\right\|_{\mathcal{P}^*} \\ \leq\;& C\left(\left(\|\boldsymbol{\Omega}^*\|_\infty + (\sqrt{K}+1)\|\boldsymbol{\mu}^*\|_\infty\right)\varphi_K + 2(\sqrt{K}+1)\varphi'_K\right)\sqrt{\frac{\log p + \log(e/\delta)}{n}}, \end{aligned}$$

with probability at least $1 - (18K+6)\delta$, where $C = \max(M_1C_1, M_2C_2, M_3C_3)$. Under regularity Condition 2.2.10, $\varphi_K \leq (c_1 + c_2^{1/2})K$, $\varphi'_K \leq (c_1 + c_2^{1/2})^2 K$. Let $C = C(c_1 + c_2^{1/2})$ and $C' = c_1^2 + c_1c_2^{1/2} + 2(c_1 + c_2^{1/2})^2$. Consequently, the upper bound for statistical error can be written as:

$$\|\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})\|_{\mathcal{P}^*} \leq \left(CK\|\boldsymbol{\Omega}^*\|_\infty + C'K^{1.5}\right)\sqrt{\frac{\log p + \log(e/\delta)}{n}},$$

with probability at least $1 - (18K+6)\delta$. ■

For the second part of Lemma 2.6.1, we are aiming to bound the statistical error arising from the estimation for diagonal term. The definition of $\mathcal{G}$ in (2.2.1) implies that $[\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})]_{\mathcal{G}}$ is a Kp-dimensional vector. Following the same derivation before, it suffices to have:

$$\begin{aligned} & \left\|[\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})]_{\mathcal{G}}\right\|_2 \\ \leq & \sqrt{Kp}\left\|[\nabla Q_n(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta})]_{\mathcal{G}}\right\|_{\max} \\ \overset{(a)}{\leq} & \sqrt{Kp}\cdot C_2\left(2\|\boldsymbol{\mu}^*\|_\infty\varphi_K + \varphi'_K\right)\sqrt{\frac{\log p + \log(e/\delta)}{n}} \\ = & \sqrt{K}\cdot C_2\left(2\|\boldsymbol{\mu}^*\|_\infty\varphi_K + \varphi'_K\right)\sqrt{\frac{p(\log p + \log(e/\delta))}{n}}, \end{aligned}$$

with probability at least $1 - (8K^2 + 2K)\delta$ where (a) comes from (2.6.36). Now combining two parts together, we end the proof of Lemma 2.6.1. ■

2.6.5 Proof of Lemma 2.5.2

For any $\boldsymbol{\Theta} \in \mathcal{M}$,

$$\begin{aligned} \frac{\mathcal{P}(\boldsymbol{\Theta})}{\|\boldsymbol{\Theta}\|_2} &= \frac{\mathcal{P}_1(\boldsymbol{\Theta})}{\|\boldsymbol{\Theta}\|_2} + \frac{\mathcal{P}_2(\boldsymbol{\Theta})}{\|\boldsymbol{\Theta}\|_2} + \frac{\mathcal{P}_3(\boldsymbol{\Theta})}{\|\boldsymbol{\Theta}\|_2} \\ &\leq \frac{M_1\sum_{k=1}^K\sum_{j=1}^p|\mu_{kj}|}{\sqrt{\sum_{k=1}^K\|\boldsymbol{\mu}_k\|_2^2}} + \frac{M_2\sum_{k=1}^K\sum_{i\neq j}|\omega_{kij}|}{\sqrt{\sum_{k=1}^K\|\boldsymbol{\Omega}_k\|_F^2}} + \frac{\sum_{i\neq j}M_3(\sum_{k=1}^K\omega_{kij}^2)^{1/2}}{\sqrt{\sum_{k=1}^K\|\boldsymbol{\Omega}_k\|_F^2}}. \end{aligned}$$

By Cauchy's inequality, we can have

$$\frac{\mathcal{P}(\boldsymbol{\Theta}_{\mathcal{M}})}{\|\boldsymbol{\Theta}_{\mathcal{M}}\|_2} \leq M_1\sqrt{Kd} + M_2\sqrt{Ks} + M_3\sqrt{s}.$$

Recall that d and s are the sparse parameter for a single cluster mean and precision matrix, respectively. This ends the proof of Lemma 2.5.2. ■

2.6.6 Proof of Lemma 2.5.4

First we consider each $\boldsymbol{\Theta}_k = \{\mu_k, \boldsymbol{\Omega}_k\}$ individually. That means we prove the following part first:

$$Q_n(\boldsymbol{\Theta}_k^{(1)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}) - \left\langle \nabla_{\boldsymbol{\Theta}_k} Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}), \boldsymbol{\Theta}_k^{(1)} - \boldsymbol{\Theta}_k^{(2)} \right\rangle \leq 0,$$

where $Q_n(\boldsymbol{\Theta}_k|\boldsymbol{\Theta})$ means we set $\boldsymbol{\Theta}_i$ $i \neq k$ to zero.

Following the same technique we use in the proof of Lemma (2.2.5), the decomposition can be made as below:

$$Q_n(\boldsymbol{\Theta}_k^{(1)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}) - \left\langle \nabla_{\boldsymbol{\Theta}_k} Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}), \boldsymbol{\Theta}_k^{(1)} - \boldsymbol{\Theta}_k^{(2)} \right\rangle = I + II,$$

where

$$\begin{aligned}
I &= \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\left\{h(\boldsymbol{\mu}_k^{(2)}, \boldsymbol{\Omega}_k^{(2)}) - h(\boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Omega}_k^{(2)})\right\}\right] \\
&-(\boldsymbol{\mu}_k^{(1)} - \boldsymbol{\mu}_k^{(2)})^\top \nabla_{\boldsymbol{\mu}_k} Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}), \\
II &= \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\{\frac{1}{2}\log\det(\boldsymbol{\Omega}_k^{(1)}) - \frac{1}{2}\log\det(\boldsymbol{\Omega}_k^{(2)})\right. \\
&\left.+h(\boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Omega}_k^{(2)}) - h(\boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Omega}_k^{(1)})\}\right] - [\mathrm{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})]^\top \nabla_{\boldsymbol{\Omega}_k} Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}).
\end{aligned}$$

Bounding I: By a little algebra, we can show that

$$I = -\frac{1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{\mu}_k^{(1)} - \boldsymbol{\mu}_k^{(2)})^\top \boldsymbol{\Omega}_k^{(2)}(\boldsymbol{\mu}_k^{(1)} - \boldsymbol{\mu}_k^{(2)}).$$

Plugging in $(\boldsymbol{\Theta}^{(t)}, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$, we have

$$I = -\frac{(1-t^*)^2}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)(\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^*)^\top \left(t^*\boldsymbol{\Omega}_k^{(t)} + (1-t^*)\boldsymbol{\Omega}_k^*\right)(\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^*).$$

Recall that $\boldsymbol{\Theta}^{(t)}$ is the solution of the optimization problem (2.5.5). The algorithm guarantees that $\boldsymbol{\Omega}_k^{(t)}$ is positive definite. Thus, from the positive definiteness of $\boldsymbol{\Omega}_k^{(t)}$ and $\boldsymbol{\Omega}_k^*$, it is sufficient to show that

$$I \leq 0 \quad \text{holds a.s..} \tag{2.6.37}$$

When plugging in $(\boldsymbol{\Theta}^*, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$, we have the same conclusion.

Bounding II: Define

$$g(\boldsymbol{\Omega}_k^{(2)}) := \frac{1}{n}\sum_{i=1}^{n}\left[L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\left\{\frac{1}{2}\log\det(\boldsymbol{\Omega}_k^{(2)}) - h\left(\boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Omega}_k^{(2)}\right)\right\}\right].$$

We rewrite II as

$$g(\boldsymbol{\Omega}_k^{(1)}) - g(\boldsymbol{\Omega}_k^{(2)}) - \left\langle \operatorname{vec}\left(\nabla g(\boldsymbol{\Omega}_k^{(2)})\right), \operatorname{vec}\left(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)}\right)\right\rangle.$$

According to Taylor expansion, we can expand $g(\boldsymbol{\Omega}_k^{(1)})$ around $\boldsymbol{\Omega}_k^{(2)}$ and obtain

$$\begin{aligned} g(\boldsymbol{\Omega}_k^{(1)}) &= g(\boldsymbol{\Omega}_k^{(2)}) + \left\langle \operatorname{vec}(\nabla g(\boldsymbol{\Omega}_k^{(2)}), \operatorname{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})\right\rangle \\ &\quad + \frac{1}{2}\left[\operatorname{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})\right]^\top \nabla^2 g(\mathbf{Z})\left[\operatorname{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})\right], \end{aligned}$$

where $\mathbf{Z} = t\boldsymbol{\Omega}_k^{(1)} + (1-t)\boldsymbol{\Omega}_k^{(2)}$ with $t \in [0,1]$. So an equivalent expression for II is given below:

$$II = \frac{1}{2}\left[\operatorname{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})\right]^\top \nabla^2 g(\mathbf{Z})\left[\operatorname{vec}(\boldsymbol{\Omega}_k^{(1)} - \boldsymbol{\Omega}_k^{(2)})\right].$$

By the definition of function g we construct, the negative Hessian matrix of function g is

$$-\nabla^2 g(\mathbf{Z}) = \frac{1}{2n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)\mathbf{Z}^{-1} \otimes \mathbf{Z}^{-1}.$$

According to the analysis in the proof of Lemma 2.2.5, $\sigma_{\min}\left(\mathbf{Z}^{-1} \otimes \mathbf{Z}^{-1}\right) = \left[\sigma_{\min}(\mathbf{Z}^{-1})\right]^2 \geq 0$. Therefore, $\nabla^2 g(\mathbf{Z})$ is a negative semi-definite matrix, which implies that $II \leq 0$ holds a.s. for any pair of points $(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)})$. Incorporating with the fact that $I < 0$, it implies that

$$Q_n(\boldsymbol{\Theta}_k^{(1)}|\boldsymbol{\Theta}^{(t-1)}) - Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}) - \left\langle \nabla_{\boldsymbol{\Theta}_k} Q_n(\boldsymbol{\Theta}_k^{(2)}|\boldsymbol{\Theta}^{(t-1)}), \boldsymbol{\Theta}_k^{(1)} - \boldsymbol{\Theta}_k^{(2)}\right\rangle \leq 0,$$

holds a.s. for pair points $(\boldsymbol{\Theta}^{(t)}, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$, $(\boldsymbol{\Theta}^{(t)}, t^*\boldsymbol{\Theta}^{(t)} + (1-t^*)\boldsymbol{\Theta}^*)$. After doing the summation from 1 to K, we finish the proof of Lemma 2.5.4. ■

2.6.7 Variable Selection Consistency

Theorem 2 . Denote the final precision matrix estimator as $\widetilde{\mathbf{\Omega}}_k$ and the set of its nonzero off-diagonal elements as $\widetilde{\mathcal{V}}_k$. Under minimal signal condition, we have, with probability tending to 1, $\widetilde{\mathcal{V}}_k = \mathcal{V}_k$ for any $k = 1, \ldots, K$.

Proof: We prove it in two steps. In Step 1, we show that $\widetilde{\mathcal{V}_k} \supset \mathcal{V}_k$, and in Step 2, we show that $\widetilde{\mathcal{V}_k} \subset \mathcal{V}_k$, both with high probability.

Step 1: In order to prove $\widetilde{\mathcal{V}_k} \supset \mathcal{V}_k$, it is sufficient to show that for any $(i, j) \in \mathcal{V}_k$ with any $k = 1, \ldots, K$, $\widetilde{\omega}_{kij} \neq 0$. Note that

$$|\omega_{kij}^{(T)}| \geq |\omega_{kij}^*| - |\omega_{kij}^{(T)} - \omega_{kij}^*| \geq |\omega_{kij}^*| - \sqrt{\sum_{i,j}(\omega_{kij}^{(T)} - \omega_{kij}^*)^2},$$

Moreover,

$$\sqrt{\sum_{i,j}(\omega_{kij}^{(T)} - \omega_{kij}^*)^2} \leq \|\mathbf{\Theta}^{(T)} - \mathbf{\Theta}^*\|_2. \tag{2.6.38}$$

According to Corollary 2.2.12 and minimal signal condition we have

$$|\omega_{kij}^{(T)}| > r_n.$$

Therefore, we see that $\widetilde{\omega}_{kij} \neq 0$, which implies $\widetilde{\mathcal{V}_k} \supset \mathcal{V}_k$.

Step 2: In order to show $\widetilde{\mathcal{V}_k} \subset \mathcal{V}_k$, we need to check that, for any $(i, j) \in \mathcal{V}_k^c$, the estimator $\widetilde{\omega}_{kij} = 0$. Note that, the estimator before the thresholding step satisfies,

$$|\omega_{kij}^{(T)}| = |\omega_{kij}^{(T)} - \omega_{kij}^*| \leq \sqrt{\sum_{i,j}(\omega_{kij}^{(T)} - \omega_{kij}^*)^2}.$$

From (2.6.38), it is known that $|\omega_{kij}^{(T)}| \leq r_n$. Therefore, the thresholding step will set $\widetilde{\omega}_{kij} = \omega_{kij}^{(T)} \mathbb{1}\{|\widehat{\omega}_{kij}| > r_n\} = 0$ with high probability. This ends the proof of Theorem 2. ■

2.7 Updates Steps of SCAN Algorithm

2.7.1 Proof of Lemma 2.1.2:

The KKT conditions for μ_{kj} to be a maximizer of $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}) - \mathcal{R}(\boldsymbol{\Theta})$ are

$$\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}\Big(\sum_{l=1}^{p}(x_{il}-\mu_{kl})\omega_{klj}\Big) = \lambda_1 \text{sign}(\mu_{kj}), \text{ when } \mu_{kj}\neq 0,$$
$$\left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}\Big(\sum_{l=1,l\neq j}^{p}(x_{il}-\mu_{kl})\omega_{klj} + x_{ij}\omega_{kjj}\Big)\right| \leq \lambda_1, \text{ when } \mu_{kj}=0.$$

Therefore, the update of $\mu_{kj}^{(t)}$ is given as:

$$\text{If } \left|\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big(\sum_{l=1,l\neq j}^{p}(x_{il}-\mu_{kl}^{(t-1)})\omega_{klj}^{(t-1)} + x_{ij}\omega_{kjj}^{(t-1)}\Big)\right| \leq \lambda_1,$$

then $\mu_{kj}^{(t)} = 0$; Else

$$\mu_{kj}^{(t)} = \Big(\omega_{kjj}^{(t-1)}\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big)^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big(\sum_{l=1}^{p} x_{il}\omega_{klj}^{(t-1)}\Big) -$$
$$\Big(\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)\Big)\Big(\sum_{l=1}^{p}\mu_{kl}^{(t-1)}\omega_{klj}^{(t-1)} - \mu_{kj}^{(t-1)}\omega_{kjj}^{(t-1)}\Big) - \lambda_1 \text{sign}(\mu_{kj}^{(t-1)})\Big\}$$

Using the definitions of $g_{1,j}(\boldsymbol{x};\boldsymbol{\Theta}_k^{(t-1)})$ and $g_{2,j}(\boldsymbol{x}_i;\boldsymbol{\Theta}_k^{(t-1)})$, we finish the proof of Lemma 2.1.2. ■

2.7.2 Proof of Lemma 2.1.3:

Recall that in (2.1.7)

$$Q_n(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)[\log\pi_k + \log f_k(\boldsymbol{x}_i;\boldsymbol{\Theta}_k)] - \mathcal{R}(\boldsymbol{\Theta}),$$

Then,

$$
\begin{aligned}
& \max_{\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)[\log \pi_k + \log f_k(\boldsymbol{x}_i;\boldsymbol{\Theta}_k)] - \mathcal{R}(\boldsymbol{\Theta}) \\
= & \max_{\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)[\log \pi_k - \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Omega}_k) \\
& -\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)] - \frac{1}{2}\mathcal{R}(\boldsymbol{\Theta}) \\
= & \max_{\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K} \frac{1}{n}\sum_{k=1}^{K}\left\{\frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)[\log\det(\boldsymbol{\Omega}_k) - (\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]\right\} - \mathcal{R}(\boldsymbol{\Theta}) \\
= & \max_{\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_K} \frac{1}{n}\sum_{k=1}^{K} n_k[\log\det(\boldsymbol{\Omega}_k) - \mathrm{trace}(\widetilde{S}_k\boldsymbol{\Omega}_k)] - \mathcal{R}(\boldsymbol{\Theta}),
\end{aligned}
$$

where the last equality is because

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k) \\
= & \frac{1}{n}\sum_{\boldsymbol{x}_i\in\mathcal{A}_k} \mathrm{trace}((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k) \\
= & \frac{1}{n}\mathrm{trace}\Big(\sum_{\boldsymbol{x}_i\in\mathcal{A}_k}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Omega}_k\Big).
\end{aligned}
$$

Then plugging in the last update of $\boldsymbol{\mu}_k$ leads to the desirable result. ■

2.8 Supporting Lemma

Lemma 2.8.1. Consider a finite number of independent centered sub-gaussian random variables X_i. Then $\sum_i X_i$ is also a centered sub-gaussian random variable. Moreover,

$$
\left\|\sum_i X_i\right\|_{\psi_2}^2 \le C\sum_i \|X_i\|_{\psi_2}^2,
$$

where C is an absolute constant.

Lemma 2.8.2. Let X, Y be two sub-Gaussian random variables. Then $Z = X\cdot Y$ is sub-exponential random variable. Moreover, there exits constant C such that

$$
\|Z\|_{\psi_1} \le C\,\|X\|_{\psi_2}\cdot\|Y\|_{\psi_2}. \tag{2.8.1}
$$

Lemma 2.8.3. Let X be sub-Gaussian random variable and Y be sub-exponential random variables. Then $X - \mathbb{E}[X]$ is also sub-Gaussian; $Y - \mathbb{E}[Y]$ is also sub-exponential. Moreover, we have

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq 2\|X\|_{\psi_2}, \ \|Y - \mathbb{E}[Y]\|_{\psi_1} \leq 2\|Y\|_{\psi_1}.$$

Lemma 2.8.4. Suppose $X_1, X_2, \ldots, X_n$ are n iid centered sub-Gaussian random variables with $\|X_1\|_{\psi_2} \leq K$. Then for every $t \geq 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \geq e \cdot \exp\left(-\frac{Cnt^2}{K^2}\right),$$

where C is an absolute constant.

Lemma 2.8.5. Suppose $X_1, X_2, \ldots, X_n$ are n iid centered sub-expoential random variables with $\|X_1\|_{\psi_1} \leq K$. Then for every $t \geq 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \geq 2 \cdot \exp\left(-C\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\} n\right),$$

where C is an absolute constant.

Lemma 2.8.6. Hoeffding's inequality Suppose $X_1, X_2 \ldots X_n$ are independent random variable, $a_1 \leq X_i \leq b_i$, then we can have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > \varepsilon\right) \leq 2\exp\left\{\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

Moreover, if $a_i = 0$ and $b_i = 1$, then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > \varepsilon\right) > 1 - 2e^{-2n\varepsilon^2}.$$

3. SPARSE AND LOW-RANK TENSOR ESTIMATION VIA CUBIC SKETCHINGS

3.1 Preliminary

We introduce notations and operations on the matrix. For matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J] \in \mathbb{R}^{I\times J}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L] \in \mathbb{R}^{K\times L}$, the *Kronecker product* is defined as a (IK)-by-(JL) matrix $\boldsymbol{A}\otimes\boldsymbol{B} = [\boldsymbol{a}_1\otimes\boldsymbol{B}\ \cdots \boldsymbol{a}_J\otimes\boldsymbol{B}]$, where $\boldsymbol{a}_j\otimes\boldsymbol{B} = (a_{j1}\boldsymbol{B}^\top, \ldots, a_{jI}\boldsymbol{B}^\top)^\top$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns $J = L$, the *Khatri-Rao product* is defined as $\boldsymbol{A}\odot\boldsymbol{B} = [\boldsymbol{a}_1\circ\boldsymbol{b}_1, \boldsymbol{a}_2\circ\boldsymbol{b}_2, \cdots, \boldsymbol{a}_J\circ\boldsymbol{b}_J] \in \mathbb{R}^{IK\times J}$. If the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the same dimension, the *Hadamard product* is their element-wise matrix product, such that $(\boldsymbol{A}*\boldsymbol{B})_{ij} = \boldsymbol{A}_{ij}\cdot\boldsymbol{B}_{ij}$. For matrix $\boldsymbol{X} = [\boldsymbol{x}_1\cdots\boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$, we also denote the vectorization $\mathrm{vec}(\boldsymbol{X}) = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top) \in \mathbb{R}^{1\times mn}$ and column-wise ℓ_2 norms as $\mathrm{Norm}(\boldsymbol{X}) = (\|\boldsymbol{x}_1\|_2, \ldots, \|\boldsymbol{x}_n\|_2) \in \mathbb{R}^{1\times n}$.

In the end, we focus on tensor notation and relevant operations. Interested readers are referred to (31) for more details. Suppose $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\times p_3}$ is an order-3 tensor, then the (i,j,k)-th element of $\mathcal{X}$ is denoted by $[\mathcal{X}]_{ijk}$. The successive tensor multiplication with vectors $\boldsymbol{u} \in \mathbb{R}^{p_2}$, $\boldsymbol{v} \in \mathbb{R}^{p_3}$ is denoted by $\mathcal{X}\times_2\boldsymbol{u}\times_3\boldsymbol{v} = \sum_{j\in[p_2],l\in[p_3]} u_j v_l \mathcal{X}_{[:,j,l]} \in \mathbb{R}^{p_1}$. We say $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\times p_3}$ is *rank-one* if it can be written as the outer product of three vectors, i.e., $\mathcal{X} = \boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3$ or $[\mathcal{X}]_{ijk} = x_{1i}x_{2j}x_{3k}$ for all i,j,k. Here "$\circ$" represents the vector outer product. $\mathcal{X}$ is symmetric if $[\mathcal{X}]_{ijk} = [\mathcal{X}]_{ikj} = [\mathcal{X}]_{jik} = [\mathcal{X}]_{jki} = [\mathcal{X}]_{kij} = [\mathcal{X}]_{kji}$ for all i,j,k. Rank-one tensor is symmetric if and only if it can be decomposed as $\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}$ for some vector $\boldsymbol{x}$.

More generally, we may decompose a tensor as the sum of rank one tensors as follows,

$$\mathcal{X} = \sum_{k=1}^{K} \eta_k \boldsymbol{x}_{1k}\circ\boldsymbol{x}_{2k}\circ\boldsymbol{x}_{3k}, \tag{3.1.1}$$

where $\eta_k \in \mathbb{R}, \boldsymbol{x}_{1k} \in \mathbb{S}^{p_1-1}, \boldsymbol{x}_{2k} \in \mathbb{S}^{p_2-1}, \boldsymbol{x}_{3k} \in \mathbb{S}^{p_3-1}$. This is the so-called CANDECOMP/PARAFAC, or CP decomposition (31) with CP-rank being defined as the minimum number K such that (3.1.1) holds. $\{\boldsymbol{x}_{1k}\}_{k=1}^K, \{\boldsymbol{x}_{2k}\}_{k=1}^K, \{\boldsymbol{x}_{3k}\}_{k=1}^K$ are called *factors* along first, second and third mode. Note that factors are normalized as unit vectors to guarantee the uniqueness of decomposition, and $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_K\}$ plays an analogous role of singular values in matrix value decomposition here. Several tensor norms also need to be introduced. The tensor Frobenius norm and tensor spectral norm are defined respectively as

$$\|\mathcal{X}\|_F = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\sum_{k=1}^{p_3}\mathcal{X}_{ijk}^2}, \ \|\mathcal{X}\|_{op} := \sup_{\boldsymbol{u}\in\mathbb{R}^{p_1},\boldsymbol{v}\in\mathbb{R}^{p_2},\boldsymbol{w}\in\mathbb{R}^{p_3}} \frac{|\langle\mathcal{X}, \boldsymbol{u}\circ\boldsymbol{v}\circ\boldsymbol{w}\rangle|}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2\|\boldsymbol{w}\|_2}, \tag{3.1.2}$$

where $\langle\mathcal{X}, \mathcal{Y}\rangle = \sum_{i,j,k}\mathcal{X}_{ijk}\mathcal{Y}_{ijk}$. Clearly, $\|\mathcal{X}\|_F^2 = \langle\mathcal{X}, \mathcal{X}\rangle$. We also consider the following sparse tensor spectral norm,

$$\|\mathcal{X}\|_s := \sup_{\substack{\|\boldsymbol{a}\|=\|\boldsymbol{b}\|=\|\boldsymbol{c}\|=1 \\ \max\{\|\boldsymbol{a}\|_0,\|\boldsymbol{b}\|_0,\|\boldsymbol{c}\|_0\}\le s}} \big|\langle\mathcal{X}, \boldsymbol{a}\circ\boldsymbol{b}\circ\boldsymbol{c}\rangle\big|. \tag{3.1.3}$$

By definition, $\|\mathcal{X}\|_s \le \|\mathcal{X}\|_{op}$. Suppose $\mathcal{X} = \boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3$ and $\mathcal{Y} = \boldsymbol{y}_1\circ\boldsymbol{y}_2\circ\boldsymbol{y}_3$ are two rank-one tensors, then it is easy to check that $\|\mathcal{X}\|_F = \|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2\|\boldsymbol{x}_3\|_2$ and $\langle\mathcal{X}, \mathcal{Y}\rangle = (\boldsymbol{x}_1^\top\boldsymbol{y}_1)(\boldsymbol{x}_2^\top\boldsymbol{y}_2)(\boldsymbol{x}_3^\top\boldsymbol{y}_3)$.

3.2 Symmetric Tensor Estimation via Cubic Sketchings

In this section, we focus on the estimation of sparse and low-rank symmetric tensors,

$$y_i = \big\langle\mathscr{T}^*, \mathscr{X}_i\big\rangle + \epsilon_i, \quad \mathscr{X}_i = \boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i \in \mathbb{R}^{p\times p\times p}, \quad i = 1, \ldots, n, \tag{3.2.1}$$

where $\boldsymbol{x}_i$ are random vectors with i.i.d. standard normal entries. As previously discussed, the tensor parameter $\mathscr{T}^*$ often satisfies certain low-dimensional structures in practice, among which the factor-wise sparsity and low-rankness (41) commonly

appear. We thus assume $\mathscr{T}^*$ is CP rank-K for $K \ll p$ and the corresponding factors are sparse,

$$\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*, \quad \text{with} \quad \|\boldsymbol{\beta}_k^*\|_2 = 1, \|\boldsymbol{\beta}_k^*\|_0 \leq s, \forall k \in [K]. \tag{3.2.2}$$

The CP low-rankness has been widely assumed in literature for its nice scalability and simple formulation (34; 43; 49). Different from the matrix factor analysis, we do not assume the tensor factors $\boldsymbol{\beta}_k^*$ here are orthogonal. On the other hand, since the low-rank tensor estimation is NP-hard in general (83), we will introduce an incoherence condition in the forthcoming Condition 3 to ensure that the correlation among different factors $\boldsymbol{\beta}_k^*$ is not too strong. Such a condition has been used in recent literature on tensor data analysis (84), compressed sensing (85), matrix decomposition (86), and dictionary learning (87).

Based on observations $\{y_i, \mathscr{X}_i\}_{i=1}^n$, we propose to estimate $\mathscr{T}^*$ via minimizing the empirical squared loss since the close-form gradient provides computational convenience,

$$\widehat{\mathscr{T}} = \underset{\mathscr{T}}{\operatorname{argmin}} \mathcal{L}(\mathscr{T}) \quad \text{subject to } \mathscr{T} \text{ is sparse and low-rank}, \tag{3.2.3}$$

where

$$\begin{aligned} \mathcal{L}(\mathscr{T}) = \mathcal{L}\left(\eta_k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\right) &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle \mathscr{T}, \mathscr{X}_i \rangle\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \eta_k \left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k\right)^3 \right)^2. \end{aligned} \tag{3.2.4}$$

Equivalently, (3.2.3) can be written as,

$$\begin{aligned} &\min_{\eta_k, \boldsymbol{\beta}_k} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \eta_k (\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)^3 \right)^2, \\ &\text{s.t. } \|\boldsymbol{\beta}_k\|_2 = 1, \|\boldsymbol{\beta}_k\|_0 \leq s, \text{ for } k \in [K]. \end{aligned} \tag{3.2.5}$$

Clearly, (3.2.5) is a non-convex optimization problem. To solve it, we propose a two-stage method as described in the next two subsections.

3.2.1 Initialization

Due to the non-convex optimization (3.2.5), a straightforward implementation of many local search algorithms, such as gradient descent and alternating minimization, may easily get trapped into local optimums and obtain sub-optimal statistical performances. Inspired by recent advances of spectral method (e.g., EM algorithm (88), phase retrieval (89), and tensor SVD (62)), we propose to evaluate an initial estimate $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ via the method of moment and sparse tensor decomposition (a variant of high-order spectral method) in the following Steps 1 and 2, respectively. The pseudo-code is given in Algorithm 1.

Step 1: Unbiased Empirical Moment Estimator. Construct the empirical moment based estimator $\mathcal{T}_s$,

$$\begin{aligned}\mathcal{T}_s := \frac{1}{6}\Big[\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i \\ - \sum_{j=1}^{p}\Big(\boldsymbol{m}_1 \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{m}_1 \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{m}_1\Big)\Big], \qquad (3.2.6)\\ \text{where } \boldsymbol{m}_1 := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i,\ \boldsymbol{e}_j \text{ is the canonical vector.}\end{aligned}$$

As will be shown in Lemma 4, $\mathcal{T}_s$ is an unbiased estimator of $\mathscr{T}^*$. The construction of (3.2.6) is motivated by high-order Stein's identity ((90); also see Theorem 9 for a complete statement). Intuitively speaking, based on the third-order score function for a Gaussian random vector $\boldsymbol{x}$: $\mathcal{S}_3(\boldsymbol{x}) = \boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x} - \sum_{j=1}^{p}(\boldsymbol{x}\circ\boldsymbol{e}_j\circ\boldsymbol{e}_j + \boldsymbol{e}_j\circ\boldsymbol{x}\circ\boldsymbol{e}_j + \boldsymbol{e}_j\circ\boldsymbol{e}_j\circ\boldsymbol{x})$, we can construct the unbiased estimator of $\mathscr{T}^*$ by properly choosing a continuously differentiable function in high-order Stein's identity. See the proof of Lemma 4 for more details.

Step 2: Sparse Tensor Decomposition. The method of moment estimator obtained in Step 1 provides an initial estimate for tensor $\mathscr{T}^*$. Then we further obtain

good initialization for the factors $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ via truncation and alternating rank-1 power iterations (51; 91),

$$\mathcal{T}_s \approx \sum_{k=1}^{K} \eta_k^{(0)} \boldsymbol{\beta}_k^{(0)} \circ \boldsymbol{\beta}_k^{(0)} \circ \boldsymbol{\beta}_k^{(0)}.$$

Note that the tensor power iterations recover one rank-1 component per time. To identify all rank-1 components, we generate a large number of different initialization vectors at first, implement a clustering step, and choose the centroids as the estimates in the initialization stage. This scheme originally appears in tensor decomposition literature (84; 91), although our problem setting and proof techniques are very different. This procedure is also very different from the matrix setting since the rank-1 component in singular value decomposition is mutually orthogonal, but we do not enforce the exact orthogonality here for $\mathscr{T}^*$.

More specifically, we firstly choose a large integer $M \gg K$ and generate M starting vectors $\{\boldsymbol{b}_m^{(0)}\}_{m=1}^M \in \mathbb{R}^p$ through sparse SVD as described in Algorithm 3. Then for each $\boldsymbol{b}_m^{(0)}$, we apply the following truncated power update:

$$\widetilde{\boldsymbol{b}}_m^{(l+1)} = \frac{\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}}{\|\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}\|_2}, \quad \boldsymbol{b}_m^{(l+1)} = \frac{T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})}{\|T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})\|_2}, \quad l = 0, \ldots,$$

where $\times_2, \times_3$ are tensor multiplication operators defined in Section 3.1 and $T_d(\boldsymbol{x}) \in \mathbb{R}^p$ is a truncation operator that sets all but the largest d entries in absolute values to zero for any vector $\boldsymbol{x} \in \mathbb{R}^p$. It is noteworthy that the symmetry of $\mathcal{T}_s$ implies

$$\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)} = \mathcal{T}_s \times_1 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)} = \mathcal{T}_s \times_1 \boldsymbol{b}_m^{(l)} \times_2 \boldsymbol{b}_m^{(l)}.$$

This means the multiplications along different modes are the same. We run power iterations till its convergence, and denote $\boldsymbol{b}_m$ as the outcome. Finally, we apply K-means to partition $\{\boldsymbol{b}_m\}_{m=1}^M$ into K clusters, then let the centroids of the output clusters be $\{\boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$ and calculate $\eta_k^{(0)} = \mathcal{T}_s \times_1 \boldsymbol{\beta}_k^{(0)} \times_2 \boldsymbol{\beta}_k^{(0)} \times_3 \boldsymbol{\beta}_k^{(0)}$ for $k \in [K]$.

Algorithm 1 Initialization in cubic sketchings

Require: response $\{y_i\}_{i=1}^n$, sketching vector $\{\boldsymbol{x}_i\}_{i=1}^n$, truncation level d, rank K, stopping error $\epsilon = 10^{-4}$.

1: **Step 1:** Calculate the moment-based tensor $\mathcal{T}_s$ as (3.2.6).

2: **Step 2:**

3: **For** $m = 1$ **to** M

Generate $\boldsymbol{b}_m^{(0)}$ through Algorithm 3.

4: **Repeat** power update:

$$\widetilde{\boldsymbol{b}}_m^{(l+1)} = \frac{\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}}{\|\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}\|_2}, \quad \boldsymbol{b}_m^{(l+1)} = \frac{T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})}{\|T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})\|_2}, \quad l = l + 1.$$

5: **Until** $\|\boldsymbol{b}_m^{(l+1)} - \boldsymbol{b}_m^{(l)}\|_2 \leq \epsilon$.

6: **End for.**

7: Perform K-means for $\{\boldsymbol{b}_m^{(l)}\}_{m=1}^M$. Denote the centroids of K clusters by $\{\boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

8: Calculate $\eta_k^{(0)} = \mathcal{T}_s \times_1 \boldsymbol{\beta}_k^{(0)} \times_2 \boldsymbol{\beta}_k^{(0)} \times_3 \boldsymbol{\beta}_k^{(0)}, k \in [K]$.

9: **return** symmetric tensor estimator $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

3.2.2 Thresholded Gradient Descent

After obtaining a warm start in the first stage, we propose to apply the thresholding gradient descent to iteratively refine the solution to the non-convex optimization problem (3.2.5). Specifically, denote $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{p\times n}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^\top \in \mathbb{R}^K$ and $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times K}$. Recall that $\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \mathcal{L}(\mathscr{T})$, and hence let

$$\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = (\nabla_{\boldsymbol{\beta}_1}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})^\top, \ldots, \nabla_{\boldsymbol{\beta}_K}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})^\top) \in \mathbb{R}^{1\times pK},$$

be the gradient function with respect to $\boldsymbol{B}$. Based on the detailed calculation in Lemma 3.11.1, $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})$ can be written as

$$\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \frac{6}{n}[\{(\boldsymbol{B}^\top\boldsymbol{X})^\top\}^3\boldsymbol{\eta} - \boldsymbol{y}]^\top[(\{(\boldsymbol{B}^\top\boldsymbol{X})^\top\}^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{X}]^\top, \tag{3.2.7}$$

where $\{(\boldsymbol{B}^\top\boldsymbol{X})^\top\}^3$ and $\{(\boldsymbol{B}^\top\boldsymbol{X})^\top\}^2$ are entry-wise cubic and squared matrices of $(\boldsymbol{B}^\top\boldsymbol{X})^\top$. Define $\varphi_h(x)$ as the thresholding function with a level h that satisfies the following minimal assumptions:

$$|\varphi_h(x) - x| \le h, \forall x \in \mathbb{R}, \text{ and } \varphi_h(x) = 0, \text{ when } |x| \le h. \tag{3.2.8}$$

Many widely used thresholding schemes, such as hard thresholding $H_h(x) = xI_{(|x|>h)}$, soft-thresholding $S_h(x) = \text{sign}(x)\max(|x| - h, x)$, satisfy (3.2.8). With slightly abuse of notations, we further define the vector thresholding function as $\varphi_h(\boldsymbol{x}) = (\varphi_h(x_1), \ldots, \varphi_h(x_p))$, for $\boldsymbol{x} \in \mathbb{R}^p$.

The initial estimates $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{B}^{(0)}$ will be updated by thresholded gradient descent in two steps summarized in Algorithm 2. It is noteworthy that only $\boldsymbol{B}$ is updated in the Step 3, while $\boldsymbol{\eta}$ will be updated in Step 4 after the update of $\boldsymbol{B}$ is finished.

Step 3: Updating $\boldsymbol{B}$ via Thresholded Gradient descent. We update $\boldsymbol{B}^{(t)}$ in each iteration step via thresholded gradient descent,

$$\text{vec}(\boldsymbol{B}^{(t+1)}) = \varphi_{\frac{\mu\boldsymbol{h}(\boldsymbol{B}^{(t)})}{\phi}}(\text{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}^{(t)}, \boldsymbol{\eta}^{(0)})). \tag{3.2.9}$$

Here,

- μ is the step size and $\phi = \sum_{i=1}^{n} y_i^2/n$ serves as an approximation for $(\sum_{k=1}^{K} \eta_k^*)^2$ (see Lemma 15);

- $\boldsymbol{h}(\boldsymbol{B}) \in \mathbb{R}^{1\times K}$ is the thresholding level defined as

$$\boldsymbol{h}(\boldsymbol{B}) = \sqrt{\frac{4\log np}{n^2}}[\{\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^3 \boldsymbol{\eta}^{(0)} - \boldsymbol{y}\}^2]^\top \{\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^2 \odot \boldsymbol{\eta}^{(0)\top}\}^2.$$

Step 4: Updating $\boldsymbol{\eta}$ via Normalization. We normalize each column of $\boldsymbol{B}^{(T)}$ and estimate the weight parameter as

$$\begin{aligned} \widehat{\boldsymbol{B}} &= (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)^\top = \Big(\frac{\boldsymbol{\beta}_1^{(T)}}{\|\boldsymbol{\beta}_1^{(T)}\|_2}, \ldots, \frac{\boldsymbol{\beta}_K^{(T)}}{\|\boldsymbol{\beta}_K^{(T)}\|_2}\Big), \\ \widehat{\boldsymbol{\eta}} &= (\widehat{\eta}_1, \ldots, \widehat{\eta}_K)^\top = \Big(\eta_1^{(0)}\|\boldsymbol{\beta}_1^{(T)}\|_2^3, \ldots, \eta_K^{(0)}\|\boldsymbol{\beta}_K^{(T)}\|_2^3\Big)^\top. \end{aligned} \tag{3.2.10}$$

The final estimator for $\mathscr{T}^*$ is

$$\widehat{\mathscr{T}} = \sum_{k=1}^{K} \widehat{\eta}_k \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k.$$

Remark 1 (Stochastic Thresholded Gradient descent). Evaluating the gradient (3.2.7) at each iteration requires $\mathcal{O}(npK^2)$ operations, which is an issue when n or p is large. To economize the computational cost, a stochastic version of thresholded gradient descent algorithm can be easily carried out by sampling a subset of summand functions (3.2.7) at each iteration. This will accelerate the procedure especially in the case of large-scale settings. Details could refer to Section 3.11.2 in the supplementary materials.

3.3 Theoretical Analysis

In this section, we establish the geometric convergence rate in optimization error and minimax optimal rate in statistical error of the proposed symmetric tensor estimator.

Algorithm 2 Thresholded gradient descent in cubic sketchings

Require: response $\{y_i\}_{i=1}^n$, sketching vector $\{\boldsymbol{x}_i\}_{i=1}^n$, step size μ, rank K, stopping error $\epsilon = 10^{-4}$, warm-start $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

Step 3: Let $t = 0$.

Repeat thresholded gradient descent

- Compute thresholding level $\boldsymbol{h}(\boldsymbol{B})$.
- Calculate the thresholded gradient descent update

$$\mathrm{vec}(\boldsymbol{B}^{(t+1)}) = \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B})}{\phi}}\Big(\mathrm{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}^{(t)}, \boldsymbol{\eta}^{(0)})\Big),$$

where $\phi = \frac{1}{n}\sum_{i=1}^n y_i^2$. The detailed form of $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}^{(0)})$ refers to (3.2.7).

Until $\|\boldsymbol{B}^{(T+1)} - \boldsymbol{B}^{(T)}\|_F \le \epsilon$.

Step 4: Perform column-wise normalization and update the weight as (3.2.10). Construct the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \widehat{\eta}_k \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k$.

return symmetric tensor estimator $\widehat{\mathscr{T}}$.

Algorithm 3 Sparse SVD

Require: tensor $\mathcal{T}_s$, cardinality parameter d.

Compute $\widetilde{\boldsymbol{\theta}} = T_d(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim \mathcal{N}(0, I_d)$.

Calculate $\boldsymbol{u}$ as the leading singular vector of $\mathcal{T}_s \times_1 \widetilde{\boldsymbol{\theta}}$.

return the sparse vector $T_d(\boldsymbol{u})/\|\boldsymbol{u}\|_2$.

3.3.1 Assumptions

Conditions 1-3 are on the true tensor parameter $\mathscr{T}^*$ while Conditions 4-5 are on the measurement scheme. The first condition guarantees the model identifiability for CP-decomposition.

Condition 1 (Uniqueness of CP-decomposition). The CP-decomposition form (3.2.2) is unique in the sense that if there exists another CP-decomposition $\mathscr{T}^* = \sum_{k=1}^{K'} \eta_k^{*'} \boldsymbol{\beta}_k^{*'} \circ \boldsymbol{\beta}_k^{*'} \circ \boldsymbol{\beta}_k^{*'}$, it must have $K = K'$ and be invariant up to a permutation of $\{1, \ldots, K\}$.

For technical purpose, we introduce the following conditions to ensure that the CP-decomposition of $\mathscr{T}^*$ has a regular form in the sense that the operator norm of $\mathscr{T}^*$ can be bounded by the largest factor and all factors are in the same order. Similar assumptions were previously used in literature (e.g., (32; 51)).

Condition 2 (Parameter space). The CP-decomposition $\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$ satisfies

$$\|\mathscr{T}^*\|_{op} \leq C\eta_{\max}^*, \quad K = \mathcal{O}(s), \quad \text{and} \quad R = \eta_{\max}^*/\eta_{\min}^* \leq C' \tag{3.3.1}$$

for some absolute constants C, C', where $\eta_{\min}^* = \min_k \eta_k^*$ and $\eta_{\max}^* = \max_k \eta_k^*$. Recall that s is the sparsity for $\boldsymbol{\beta}_k^*$.

The performance of Step 2, i.e. the tensor decomposition for initialization, is crucial to the final estimation. However, as shown in the seminal work of (83), the estimation of the low-rank tensor is NP-hard in general. Hence, we impose the following incoherence condition that is widely used in tensor decomposition literature (51; 91).

Condition 3 (Parameter incoherence). The true tensor components are incoherent such that

$$\Gamma := \max_{1 \leq k_1 \neq k_2 \leq K} |\langle \boldsymbol{\beta}_{k_1}^*, \boldsymbol{\beta}_{k_2}^* \rangle| \leq \min\{C'' K^{-\frac{3}{4}} R^{-1}, s^{-\frac{1}{2}}\},$$

where R is the singular value ratio defined in (3.3.1) and C'' is some small constant.

Remark 3.3.1. The preceding incoherence condition has been widely used in different scenarios in recent high-dimensional research, such as compressed sensing (85), matrix decomposition (86), and dictionary learning (87). It can also be viewed as a relaxation of orthogonality: if $\{\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*\}$ are mutually orthogonal, Γ equals zero. In addition, we can show from both theory (Lemma 28 in the supplementary materials) and simulation (Section 3.6) that the low-rank tensor $\mathscr{T}^*$ induced by (3.2.2) satisfies the incoherence condition with high probability, if the component vectors $\boldsymbol{\beta}_k^*$ are randomly generated, say from Gaussian distribution.

We also introduce the following conditions on noise and sample complexity.

Condition 4 (Sub-exponential noise). The noise $\{\epsilon_i\}_{i=1}^n$ are i.i.d. randomly generated with mean 0 and variance σ^2 satisfying $0 < \sigma < C\sum_{k=1}^K \eta_k^*$. (ϵ_i/σ) is sub-exponential distributed, i.e., there exists constant $C_\epsilon > 0$ such that $\|(\epsilon_i/\sigma)\|_{\psi_1} := \sup_{p\geq 1} p^{-1}(\mathbb{E}|\epsilon_i/\sigma|^p)^{1/p} \leq C_\epsilon$, and independent of $\{\mathscr{X}_i\}_{i=1}^n$.

The sample complexity condition is crucial for our algorithm, especially in the initialization stage. Ignoring any polylog factors, Condition 5 is even weaker than the sparse matrix estimation case ($n \gtrsim s^2$) in (89).

Condition 5 (Sample complexity). We assume a sufficient number of observations is observed,

$$n \geq C''' K^2 (s\log(ep/s))^{\frac{3}{2}} \log^4 n.$$

3.3.2 Main Theoretical Results

Our main Theorem 3 shows that based on a good initializer, the output from the proposed thresholded gradient descent can achieve optimal statistical rate after sufficient iterations. Here, we define a contraction parameter

$$0 < \kappa = 1 - 32\mu K^{-2} R^{-\frac{8}{3}} < 1,$$

and also denote $\mathcal{E}_1 = 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2$ and $\mathcal{E}_2 = C_0\eta_{\min}^{*-\frac{4}{3}}/16$ for some $C_0 > 0$.

Theorem 3 (Statistical Error and Optimization Error). Suppose Conditions 3-5 hold and the initial estimator $\{\boldsymbol{\beta}_k^{(0)}, \eta_k^{(0)}\}_{k=1}^K$ satisfies

$$\max_{1\le k\le K}\left\{\left\|\boldsymbol{\beta}_k^{(0)}-\boldsymbol{\beta}_k^*\right\|_2, |\eta_k^{(0)}-\eta_k^*|\right\} \lesssim K^{-1}, \tag{3.3.2}$$

with probability at least $1-\mathcal{O}(1/n)$ and $|\mathrm{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$. Assume the step size $\mu \le \mu_0$, where μ_0 is defined in (3.9.6). Then, the output from the thresholded gradient descent update in (3.2.9) satisfies:

- For any $t = 0, 1, 2, \ldots$, the factor-wise estimator satisfies

$$\sum_{k=1}^K \left\| \sqrt[3]{\eta_k^{(0)}}\boldsymbol{\beta}_k^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \le \mathcal{E}_1\kappa^t + \mathcal{E}_2\frac{\sigma^2 s\log p}{n}, \tag{3.3.3}$$

  with probability at least $1-\mathcal{O}(tKs/n)$.

- When the total number of iterations is no smaller than

$$T^* = \left(\log(\frac{n}{\sigma^2 s\log p}\vee 1) + \log\frac{\mathcal{E}_1}{\mathcal{E}_2}\right)/\log\kappa^{-1}, \tag{3.3.4}$$

  there exists a constant C_1 (independent of K, s, p, n, σ^2) s.t. the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \eta_k^{(0)}\boldsymbol{\beta}_k^{(T^*)}\circ\boldsymbol{\beta}_k^{(T^*)}\circ\boldsymbol{\beta}_k^{(T^*)}$ is upper bounded by

$$\left\|\widehat{\mathscr{T}}-\mathscr{T}^*\right\|_F^2 \le \frac{C_1\sigma^2 Ks\log p}{n}, \tag{3.3.5}$$

  with probability at least $1-\mathcal{O}(T^*Ks/n)$.

Remark 2 . From (3.3.3), the error bound can be decomposed into an optimization error $\mathcal{E}_1\kappa^t$ (which decays with a geometric rate as iterations) and a statistical error $\mathcal{E}_2\frac{\sigma^2 s\log p}{n}$ (which does not decay as iterations). In particular, the convergence rate of the optimization error relies on the rank K and the singular value ratio R in the sense that the smaller K or R, the faster convergence. Also from (3.3.5), we note that in the special case that $\sigma = 0$, $\widehat{\mathscr{T}}$ exactly recover $\mathscr{T}^*$ with high probability.

The next theorem shows that Steps 1 and 2 of Algorithm 1 provides a good initializer required in Theorem 3.

Theorem 4 (Initialization Error). Suppose the number of initializations $L \geq K^{C_3\gamma^{-4}}$, where γ is a constant defined in (3.9.3). Given that Conditions 1-4 hold, the initial estimator obtained from Steps 1-2 with a truncation level $s \leq d \leq Cs$ satisfies

$$\max_{1\leq k\leq K}\left\{\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \leq C_2 K R \delta_{n,p,s} + \sqrt{K}\Gamma^2, \tag{3.3.6}$$

and $|\mathrm{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$ with probability at least $1 - 5/n$, where

$$\delta_{n,p,s} = (\log n)^3\Big(\sqrt{\frac{s^3\log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}}\Big). \tag{3.3.7}$$

Moreover, if the sample complexity condition 5 is satisfied, then the above bound satisfies (3.3.2).

Remark 3 (Interpretation of initialization error). The upper bound of (3.3.6) consists of two terms, which corresponds to the approximation error of $\mathcal{T}_s$ to $\mathscr{T}^*$ and the incoherence condition of $\boldsymbol{\beta}_k^*$'s, respectively. Especially, the former converges to zero as n grows while the latter does not. This indicates that the convergence rate of the initial estimate is significantly slower than that of the final estimate after iterative updates, unless

$$n \gtrsim (s\log(ep/s))^2 \quad \text{and} \quad \Gamma^2 \lesssim \sqrt{\frac{s\log(ep/s)}{nK}}.$$

More detailed numerical comparisons will be provided later in Section 3.6.

The proof of Theorems 3 and 4 are involved and postponed to Section 3.9.1-3.9.2 in the supplementary materials. The combination of Theorems 3 and 4 immediately yields the following upper bound for the final estimate as one main result in this paper.

Theorem 5 (Upper Bound). Suppose Conditions 1 – 5 hold, $s \leq d \leq Cs$. After T^* iterations, there exists a constant C_1 not depending on K, s, p, n, σ^2, such that the proposed procedure yields

$$\Big\|\widehat{\mathscr{T}} - \mathscr{T}^*\Big\|_F^2 \leq \frac{C_1\sigma^2 K s\log p}{n}, \tag{3.3.8}$$

with probability at least $1 - \mathcal{O}(T^*Ks/n)$, where T^* is defined in (3.3.4).

The above upper bound turns out to match with the minimax lower bound for a large class of sparse and low rank tensors.

Theorem 6 (Lower Bound). Consider the following class of sparse and low-rank tensors,

$$\mathcal{F}_{p,K,s} = \left\{ \mathscr{T} : \begin{array}{l} \mathscr{T} = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \|\boldsymbol{\beta}_k\|_0 \le s, \text{ for } k \in [K], \\ \mathscr{T} \text{ satisfies Conditions 1, 2, and 3.} \end{array} \right\}. \tag{3.3.9}$$

Suppose that $\{\mathscr{X}_i\}_{i=1}^n$ are i.i.d standard normal cubic sketchings with i.i.d. $N(0, \sigma^2)$ noise in (3.2.1). We have the following lower bound result,

$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}_{p,K,s}} \mathbb{E} \left\| \widetilde{\mathscr{T}} - \mathscr{T} \right\|_F^2 \ge c\sigma^2 \frac{Ks \log(ep/s)}{n}.$$

The proof of Theorem 6 is deferred to Section 3.9.3 in the supplementary materials. Combining Theorem 5 and Theorem 6 together, we immediately obtain the following minimax-optimal rate for sparse and low-rank tensor estimation with cubic sketchings when $\log p \asymp \log(p/s)$:

$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T}^* \in \mathcal{F}_{p,K,s}} \mathbb{E} \left\| \widetilde{\mathscr{T}} - \mathscr{T}^* \right\|_F^2 \asymp \sigma^2 \frac{Ks \log(p/s)}{n}. \tag{3.3.10}$$

The rate in (3.3.10) sheds light upon the effect of dimension p, noise level σ^2, sparsity s, sample size n and rank K to the estimation performance.

Remark 4 (Non-sparse low-rank tensor estimation via cubic-sketchings). When the low-rank tensor $\mathscr{T}^*$ is not necessarily sparse, i.e.,

$$\mathscr{T}^* \in \mathcal{F}_{p,K} = \left\{ \mathscr{T} : \begin{array}{l} \mathscr{T} = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \text{ for } k \in [K], \\ \mathscr{T} \text{ satisfies Conditions 1, 2, and 3} \end{array} \right\},$$

we can apply the proposed procedure with all the truncation/thresholding steps removed. If $n \ge \mathcal{O}(p^{3/2})$, one can apply similar arguments of Theorems 3-5 to show that the output estimation $\widehat{\mathscr{T}'}$ satisfies

$$\left\| \widehat{\mathscr{T}'} - \mathscr{T}^* \right\|_F^2 \lesssim \frac{\sigma^2 Kp}{n}. \tag{3.3.11}$$

for any $\mathscr{T}^* \in \mathcal{F}_{p,K}$ with high probability. Furthermore, similar arguments of Theorem 6 imply that the rate in (3.3.11) is minimax optimal.

Remark 5 (Comparison with existing matrix results). Our cubic sketching tensor results are not a direct extension of the existing matrix results. For example, (55; 56) studied the low-rank matrix recovery based on rank-1 projections: $y_i = \boldsymbol{x}_i^\top \boldsymbol{T} \boldsymbol{x}_i + \epsilon_i$ based on the convex nuclear norm minimization. The theoretical properties of their estimate are analyzed under a ℓ_1/ℓ_2-RIP or Restricted Uniform Boundedness (RUB) condition. However, tensor nuclear norm is computationally infeasible and following the arguments in (89; 92), one can check that our cubic sketching framework does not satisfy RIP or RUB conditions in general. Thus, these previous results cannot be directly applied.

In addition, the analysis of gradient updates for the tensor case is significantly more complicated than the matrix case. First, we require high-order concentration inequalities for the tensor case since the cubic-sketching tensor leads to high-order products of sub-Gaussian random variables (see Section 3.3.3 for details). The necessity of high-order expansions in the analysis of gradient updates for the tensor case also significantly increases the hardness of the problem. To ensure the geometric convergence, we need much more subtle controls on the regularity conditions comparing to the ones in the matrix case (92).

3.3.3 Key Lemmas: High-order Concentration Inequalities

As mentioned earlier, one major challenge for theoretical analysis of cubic sketching is to handle heavy tails of high-order Gaussian moments. One can only handle up-to second moments of sub-Gaussian random variables by directly applying the existing Hoeffding's or Bernstein's concentration inequalities. Rather, we need to develop the following two high-order concentration inequalities as technical tools: Lemma 1 characterizes the tail bounds for the sum of sub-Gaussian products, and Lemma 2 provides the concentration inequalities for Gaussian cubic sketchings. The proofs of Lemma 1 and 2 are given in Section 3.8.2.

Lemma 1 (Concentration inequality for sum of sub-Gaussian products). Suppose $\boldsymbol{X}_i = (\boldsymbol{x}_{1i}^\top, \ldots, \boldsymbol{x}_{mi}^\top)^\top \in \mathbb{R}^{m\times p}$, $i \in [n]$ are n i.i.d random matrices. Here, $\boldsymbol{x}_{ij}$ is the j-th row of $\boldsymbol{X}_i$ and suppose it is an isotropic sub-Gaussian vector. Then for any vectors $\boldsymbol{a} = (a_1 \ldots, a_n) \in \mathbb{R}^n$, $\{\boldsymbol{\beta}_j\}_{j=1}^m \subseteq \mathbb{R}^p$, and $0 < \delta < 1$, we have

$$\begin{aligned}&\Big|\sum_{i=1}^n a_i \prod_{j=1}^m (\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) - \mathbb{E}\Big(\sum_{i=1}^n a_i \prod_{j=1}^m (\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)\Big)\Big| \\ &\le C \prod_{j=1}^m \|\boldsymbol{\beta}_j\|_2 \Big(\|\boldsymbol{a}\|_\infty (\log \delta^{-1})^{m/2} + \|\boldsymbol{a}\|_2 (\log \delta^{-1})^{1/2}\Big),\end{aligned}$$

with probability at least $1-\delta$ for some constant C.

Note that in Lemma 1, entries in each matrix $\boldsymbol{X}_i$ are not necessarily independent even $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent matrices. Building on Lemma 1, Lemma 2 provides a generic spectral-type concentration inequality that can be used to quantify the approximation error for $\mathcal{T}_s$ introduced in Step 1 of the proposed procedure.

Lemma 2 (Concentration inequality for Gaussian cubic sketchings). Suppose $\{\boldsymbol{x}_{1i}\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_{p_1})$, $\{\boldsymbol{x}_{2i}\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_{p_2})$, $\{\boldsymbol{x}_{3i}\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_{p_3})$, $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_2}$, $\boldsymbol{\beta}_3 \in \mathbb{R}^{p_3}$ are fixed vectors.

- Define $M_{\text{nsy}} = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{2i} \circ \boldsymbol{x}_{3i}, \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \boldsymbol{\beta}_3 \rangle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{2i} \circ \boldsymbol{x}_{3i}$. Then $\mathbb{E}(M_{\text{nsy}}) = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \boldsymbol{\beta}_3$, and
$$\begin{aligned}&\Big\|M_{\text{nsy}} - \mathbb{E}(M_{\text{nsy}})\Big\|_s \\ &\le C(\log n)^3 \Big(\sqrt{\frac{s^3 \log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}}\Big)\|\boldsymbol{\beta}_1\|_2 \|\boldsymbol{\beta}_2\|_2 \|\boldsymbol{\beta}_3\|_2,\end{aligned}$$
with probability at least $1 - 10/n^3$.

- Define $M_{\text{sym}} = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i}, \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \rangle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i}$. Then $\mathbb{E}(M_{\text{sym}}) = 6\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 + 3\sum_{m=1}^p (\boldsymbol{\beta}_1 \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{\beta}_1 \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ \boldsymbol{\beta}_1)$, and
$$\begin{aligned}&\Big\|M_{\text{sym}} - \mathbb{E}(M_{\text{sym}})\Big\|_s \\ &\le C(\log n)^3 \Big(\sqrt{\frac{s^3 \log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}}\Big)\|\boldsymbol{\beta}_1\|_2^3,\end{aligned}$$
with probability at least $1 - 10/n^3$.

Here, C is an absolute constant and $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (3.1.3).

Note that M_{sym} is the major term in the unbiased empirical moment estimator $\mathcal{T}_s$ in (3.2.6), while M_{nsy} corresponds to the non-symmetric unbiased empirical moment estimator $\mathcal{T}$ that will be introduced later in (3.5.4).

3.4 Application to High-Order Interaction Effect Models

In this section, we estimate high-order interaction effect models in the cubic sketching framework. Specifically, we consider the following three-way interaction model

$$y_l = \xi_0 + \sum_{i=1}^{p} \xi_i z_{li} + \sum_{i,j=1}^{p} \gamma_{ij} z_{li} z_{lj} + \sum_{i,j,k=1}^{p} \eta_{ijk} z_{li} z_{lj} z_{lk} + \epsilon_l, \quad l = 1, \ldots, n. \tag{3.4.1}$$

Here $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\eta}$ are coefficients for main effect, pairwise interaction, and triple-wise interaction, respectively. Importantly, (3.4.1) can be reformulated as the following tensor form (also see the left panel in Figure 1.2)

$$y_l = \langle \mathcal{B}, \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \rangle + \epsilon_l, \quad l = 1, \ldots, n, \tag{3.4.2}$$

where $\boldsymbol{x}_l = (1, \boldsymbol{z}_l^\top)^\top \in \mathbb{R}^{p+1}$ and $\mathcal{B} \in \mathbb{R}^{(p+1)\times(p+1)\times(p+1)}$ is a tensor parameter corresponding to coefficients in the following way:

$$\begin{cases} \mathcal{B}_{[0,0,0]} = \xi_0, \\ \mathcal{B}_{[1:p,1:p,1:p]} = (\eta_{ijk})_{1\le i,j,k\le p}, \\ \mathcal{B}_{[0,1:p,1:p]} = \mathcal{B}_{[1:p,0,1:p]} = \mathcal{B}_{[1:p,1:p,0]} = (\gamma_{ij}/3)_{1\le i,j\le p}, \\ \mathcal{B}_{[0,0,1:p]} = \mathcal{B}_{[0,1:p,0]} = \mathcal{B}_{[1:p,0,0]} = (\xi_i/3)_{1\le i\le p}. \end{cases} \tag{3.4.3}$$

We next argue that it is reasonable to assume $\mathcal{B}$ is low rank and sparse in the tensor formulation of high-order interaction models. First, in modern biomedical research such as (93), only a small portion of coefficients contribute to the response, leading to a highly sparse $\mathcal{B}$. Further, (94) suggested that for the low-enough rank it is suitable

to model sparse tensors as arising from sparse loadings, saying CP-decomposition. Moreover, this low-rank-and-sparse assumption (or approximation) seems necessary when the sample size is limited. Specifically, we assume $\mathcal{B}$ is of CP rank-K with s-sparse factors, where $K, s \ll p$. It is easy to see that the number of parameters in (3.4.4) is $K(p+1)$, which is significantly smaller than $(p+1)^3$, the total number of parameters in the original three-way interaction effect model (3.4.1). In this case, (3.4.2) can be written as

$$\begin{aligned} y_l = \Big\langle \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \Big\rangle + \epsilon_l, \quad l = 1, \ldots, n, \\ \text{where} \quad \|\boldsymbol{\beta}_k\|_2 = 1, \quad \|\boldsymbol{\beta}_k\|_0 \le s, \quad k \in [K]. \end{aligned} \tag{3.4.4}$$

By assuming $\boldsymbol{z}_l \overset{iid}{\sim} N_p(0, \boldsymbol{I}_p)$, the high-order interaction effect model (3.4.2) reduces to the symmetric tensor estimation model (3.2.1) with the only difference that the first coordinate of $\boldsymbol{x}_l$, i.e., the intercept, is always 1. To accommodate this slight difference, we only need to adjust the initial unbiased estimate in the above two-step procedure. We first obtain $\mathcal{T}_s$ in (3.2.6) by replacing $\boldsymbol{x}_i$ therein by $\boldsymbol{x}_l$, where $\boldsymbol{x}_l$ corresponds the l-th observation

$$\begin{aligned} \mathcal{T}_s =& \frac{1}{6n} \sum_{l=1}^{n} y_l \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l - \frac{1}{6} \sum_{j=1}^{p} (\boldsymbol{a} \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{a} \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{a}), \\ \text{where} \quad & \boldsymbol{a} = \frac{1}{n} \sum_{l=1}^{n} y_l \boldsymbol{x}_l, \end{aligned} \tag{3.4.5}$$

then construct empirical-moment-based initial tensor $\mathcal{T}_{s'}$ as

- For $i, j, k \neq 0$, $\mathcal{T}_{s'[i,j,k]} = \mathcal{T}_{s[i,j,k]}$. And $\mathcal{T}_{s'[i,j,0]} = \mathcal{T}_{s[i,j,0]}, \mathcal{T}_{s'[0,j,k]} = \mathcal{T}_{s[0,j,k]}, \mathcal{T}_{s'[i,0,k]} = \mathcal{T}_{s[i,0,k]}$.
- For $i \neq 0$, $\mathcal{T}_{s'[0,0,i]} = \mathcal{T}_{s'[0,i,0]} = \mathcal{T}_{s'[i,0,0]} = \frac{1}{3}\mathcal{T}_{s[0,0,i]} - \frac{1}{6}(\sum_{k=1}^{p} \mathcal{T}_{s[k,k,i]} - (p+2)a_i)$.
- $\mathcal{T}_{s'[0,0,0]} = \frac{1}{2p-2}(\sum_{k=1}^{p} \mathcal{T}_{s[0,k,k]} - (p+2)\mathcal{T}_{s[0,0,0]})$.

Lemma 5 verifies that $\mathcal{T}_{s'}$ is an unbiased estimator for $\mathcal{B}$.

Theoretical results in Section 3.3 imply the following upper and lower bound results in this particular example.

Corollary 1 . Suppose that $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are i.i.d. standard Gaussian random vectors and $\mathcal{B}$ satisfies Conditions 1, 2 and 3. The output, denoted as $\widehat{\mathcal{B}}$, from the proposed Algorithms 1 and 2 based on $\mathcal{T}_{s'}$ satisfies

$$\left\|\widehat{\mathcal{B}} - \mathcal{B}\right\|_F^2 \leq C\frac{\sigma^2 K s \log p}{n} \tag{3.4.6}$$

with high probability. On the other hand, considering the following class of $\mathcal{B}$,

$$\mathcal{F}_{p+1,K,s} = \left\{ \mathcal{B} : \begin{array}{l} \mathcal{B} = \sum_{k=1}^K \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \|\boldsymbol{\beta}_k\|_0 \leq s, \text{ for } k \in [K], \\ \mathcal{B} \text{ satisfies Conditions 1, 2, and 3,} \end{array} \right\}.$$

then the following lower bound holds,

$$\inf_{\widehat{\mathcal{B}}} \sup_{\mathcal{B} \in \mathcal{F}_{p+1,K,s}} \mathbb{E} \left\|\widehat{\mathcal{B}} - \mathcal{B}\right\|_F^2 \geq C\frac{\sigma^2 K s \log p}{n}.$$

3.5 Non-symmetric Tensor Estimation Model

In this section, we extend the previous results to the non-symmetric tensor case. Specifically, we have $\mathscr{T}^* \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, and

$$y_i = \langle \mathscr{T}^*, \mathscr{X}_i \rangle + \epsilon_i, \;\; \mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \;\; i \in [n], \tag{3.5.1}$$

where $\boldsymbol{u}_i \in \mathbb{R}^{p_1}, \boldsymbol{v}_i \in \mathbb{R}^{p_2}, \boldsymbol{w}_i \in \mathbb{R}^{p_3}$ are random vectors with i.i.d. standard normal entries. Again, we assume $\mathscr{T}^*$ is sparse and low-rank in a similar sense that

$$\begin{aligned} &\mathscr{T}^* = \sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*, \\ &\|\boldsymbol{\beta}_{1k}^*\|_2 = \|\boldsymbol{\beta}_{2k}^*\|_2 = \|\boldsymbol{\beta}_{3k}^*\|_2 = 1, \quad \max\{\|\boldsymbol{\beta}_{1k}^*\|_0, \|\boldsymbol{\beta}_{2k}^*\|_0, \|\boldsymbol{\beta}_{3k}^*\|_0\} \leq s. \end{aligned} \tag{3.5.2}$$

Denote the following:

- $\boldsymbol{B}_1 = (\boldsymbol{\beta}_{11}, \cdots, \boldsymbol{\beta}_{1K})$, $\boldsymbol{B}_2 = (\boldsymbol{\beta}_{21}, \cdots, \boldsymbol{\beta}_{2K})$, $\boldsymbol{B}_3 = (\boldsymbol{\beta}_{31}, \cdots, \boldsymbol{\beta}_{3K})$,
- $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^\top, \boldsymbol{y} = (y_1, \ldots, y_n)^\top$.

Then, the empirical risk function can be written compactly as

$$\mathcal{L}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{\eta}) = \frac{1}{n}\Big\|(\boldsymbol{U}^\top \boldsymbol{B}_1) * (\boldsymbol{V}^\top \boldsymbol{B}_2) * (\boldsymbol{W}^\top \boldsymbol{B}_3) \cdot \boldsymbol{\eta} - \boldsymbol{y}\Big\|_2^2. \tag{3.5.3}$$

We note that (3.5.3) is non-convex, but fortunately tri-convex in terms of $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ and $\boldsymbol{B}_3$. This allows us to develop a block-wise thresholded gradient descent algorithm as detailed below.

The major steps of the estimation procedure for non-symmetric tensors are sketched below. The complete algorithm is deferred to Section 3.10.1 in the supplementary materials.

Step 1: (Method of Tensor Moments) Construct an empirical-moment-based estimator,

$$\mathcal{T} := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \in \mathbb{R}^{p_1 \times p_2 \times p_3}. \tag{3.5.4}$$

to which sparse tensor decomposition is applied for initialization.

Step 2: (Block-wise Gradient descent) Lemma 17 in the supplementary materials shows that the gradient function for (3.5.3) with respect to $\boldsymbol{B}_1$ can be written as

$$\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{\eta}) = \boldsymbol{D}^\top (\boldsymbol{C}_1^\top \odot \boldsymbol{U})^\top \in \mathbb{R}^{1\times(p_1 K)}, \tag{3.5.5}$$

where $\boldsymbol{D} = (\boldsymbol{B}_1^\top \boldsymbol{U})^\top * (\boldsymbol{B}_2^\top \boldsymbol{V})^\top * (\boldsymbol{B}_3^\top \boldsymbol{W})^\top \boldsymbol{\eta} - \boldsymbol{y}$ and $\boldsymbol{C}_1 = (\boldsymbol{B}_2^\top \boldsymbol{V})^\top * (\boldsymbol{B}_3^\top \boldsymbol{W})^\top \odot \boldsymbol{\eta}^\top$. For $t = 1, \ldots, T$ we fix $\boldsymbol{B}_2^{(t)}$ and $\boldsymbol{B}_3^{(t)}$ and update $\boldsymbol{B}_1^{(t+1)}$ via block-wise thresholded gradient descent,

$$\mathrm{vec}(\boldsymbol{B}_1^{(t+1)}) = \varphi_{\frac{\mu h(\boldsymbol{B}_1^{(t)})}{\phi}}\Big(\mathrm{vec}(\boldsymbol{B}_1^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1^{(t)}, \boldsymbol{B}_2^{(t)}, \boldsymbol{B}_3^{(t)}, \boldsymbol{\eta})\Big),$$

where $\phi = \sum_{i=1}^n y_i^2/n$, μ is the step size and $\boldsymbol{h}(\boldsymbol{B}) = \sqrt{\frac{4\log np}{n^2}\{\boldsymbol{D}^2\}^\top\{\boldsymbol{C}^2\}}$. The updates of $\boldsymbol{B}_2, \boldsymbol{B}_3$ are similar.

The main theoretical analysis is different from the symmetric one in two folds. First, the non-symmetric cubic sketching tensor is formed by three independent

Gaussian vectors. This leads to differences in many high-order moment calculations. Second, the corresponding CP-decomposition, i.e., (3.5.2), essentially forms a bi-convex optimization. In this case, standard convex analysis for vanilla gradient descent (95) could be applied given a good enough initialization.

We impose similar regularity conditions whose detailed forms and explanations are postponed to Section 3.10.1 and the proof to Section 3.10.2. The main theorems for non-symmetric tensor estimation are presented as follows.

Theorem 7 (Upper Bound). Suppose Conditions 6 – 9 in the supplementary materials hold and $n \gtrsim (s\log(p_0/s))^{3/2}$, where $p_0 = \max\{p_1, p_2, p_3\}$. For any $t = 0, 1, 2, \ldots$, the estimation output by Algorithm 3.10.1 satisfies

$$\sum_{k=1}^{K}\sum_{j=1}^{3}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_{jk}^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_{jk}^*\right\|_2^2 \leq \mathcal{O}_p\Big(\kappa^t + \frac{\sigma^2 s\log p_0}{n}\Big)$$

for some $0 < \kappa < 1$. When the total number of iterations is no smaller than $\log(\frac{n}{\sigma^2 s\log p_0} \vee 1)/\log\kappa^{-1}$, the final estimator $\widehat{\mathscr{T}}$ satisfies

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \leq \mathcal{O}_p\Big(\frac{\sigma^2 K s\log p_0}{n}\Big).$$

Theorem 8 (Lower Bound). Consider the class of incoherent sparse and low-rank tensors $\mathcal{F} = \{\mathscr{T} : \mathscr{T} = \sum_{k=1}^{K}\boldsymbol{\beta}_{1k}\circ\boldsymbol{\beta}_{2k}\circ\boldsymbol{\beta}_{3k}, \|\boldsymbol{\beta}_{i,k}\|_0 \leq s \text{ for } i = 1, 2, 3, k = 1, \ldots, K\}$. If $\{\mathscr{X}_i\}_{i=1}^n$ are i.i.d standard normal cubic sketchings with i.i.d. $N(0, \sigma^2)$ noises in (3.5.1), the following lower bound holds,

$$\inf_{\widehat{\mathscr{T}}}\sup_{\mathscr{T}\in\mathcal{F}}\mathbb{E}\left\|\widehat{\mathscr{T}} - \mathscr{T}\right\|_F^2 \geq \frac{C\sigma^2 sK\log(e\cdot p_0/s)}{n}. \tag{3.5.6}$$

It can be seen from Theorems 7 and 8 that our proposed algorithm achieves minimax-optimal estimation error rate in the class of $\mathcal{F}$ as long as $\log(p_0) \asymp \log(p_0/s)$.

3.6 Numerical Results

In this section, we empirically examine the effect of noise level, CP-rank, sample size, dimension, and sparsity on the estimation performance. We also examine

the robustness of the proposed algorithm under the setting when the incoherence assumption used in theory fails to hold.

In each setting, we generated $\mathscr{T}^* = \sum_{k=1}^{K} \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$, where $|\text{supp}(\boldsymbol{\beta}_k^*)| = s$ was uniformly selected from $\{1, \ldots, p\}$, the nonzero entries of $\boldsymbol{\beta}_k^*$ were drawn from standard Gaussian distribution. Next we normalized each vector $\boldsymbol{\beta}_k^*$ and aggregated the coefficient as η_k^*. The cubic sketchings $\{\mathscr{X}_i\}_{i=1}^n$ were generated as $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$, where $\{\boldsymbol{x}_i\}_{i=1}^n$ were from standard Gaussian distribution. The noise $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma^2)$ or Laplace$(0, \sigma/\sqrt{2})$. Additionally, we adopt the following stopping rules: (1) the initialization iteration (Step 2 in Algorithm 1) is stopped if $\|\boldsymbol{b}_m^{(l+1)} - \boldsymbol{b}_m^{(l)}\|_2 \leq 10^{-6}$; (2) the gradient update iteration (Step 3 in Algorithm 2) is stopped if $\|\boldsymbol{B}^{(T+1)} - \boldsymbol{B}^{(T)}\|_F \leq 10^{-6}$. All presented results were based on 200 repetitions. The code was written in R and implemented on an Intel Xeon-E5 processor with 64 GB of RAM.

First, we consider the percent of successful recovery in the noiseless case. Let $K = 3$, $s/p = 0.3$, $p = 30$ or 50, so that the total number of unknown parameters in $\mathscr{T}^*$ is 2.7×10^4 or 1.25×10^5. The sample size n ranges from 500 to 6000. The recovery is called successful if the relative error $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F / \|\mathscr{T}^*\|_F < 10^{-4}$. We report the percent of successful recovery in Figure 3.1.

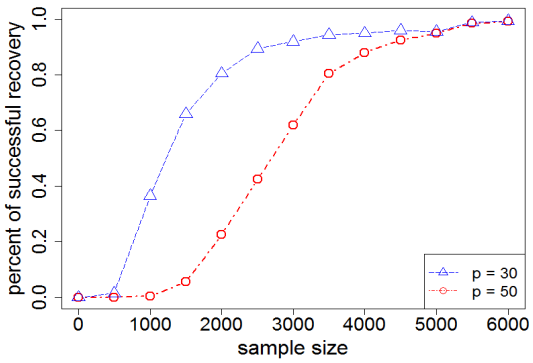

Fig. 3.1. Percent of successful recovery with varying sample size.

It is clear from Figure 3.1 that the empirical relation with dimensionality and sample size is consistent with our theory.

We then move to the noisy case where the empirical estimation error is examined. We select $K = 3$, $s/p = 0.3$, $p = 30$ or 50, $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma^2)$ and consider two specific scenarios: (1) sample size $n = 6000$, 8000, or 10000, $s/p = 0.3$, the noise level σ varies from 0 to 200; (2) noise level $\sigma = 200$, sample size n varies from 4000 to 10000, $p = 30$, $s/p = 0.1, 0.3, 0.5$. The estimation errors in terms of $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F$ under these two scenarios are plotted in Figures 3.2 and 3.3, respectively. From these results, we can see that the proposed algorithm achieves reasonable estimation performance: Algorithms 1 and 2 yield more accurate estimation with smaller variance σ^2 and/or large value of sample size n.

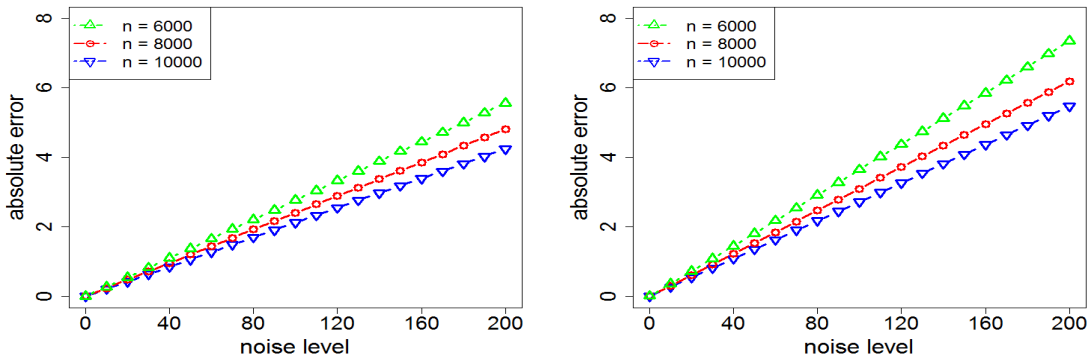

Fig. 3.2. Estimation error for different noise levels. The left panel is $p = 30$ and the right panel is $p = 50$.

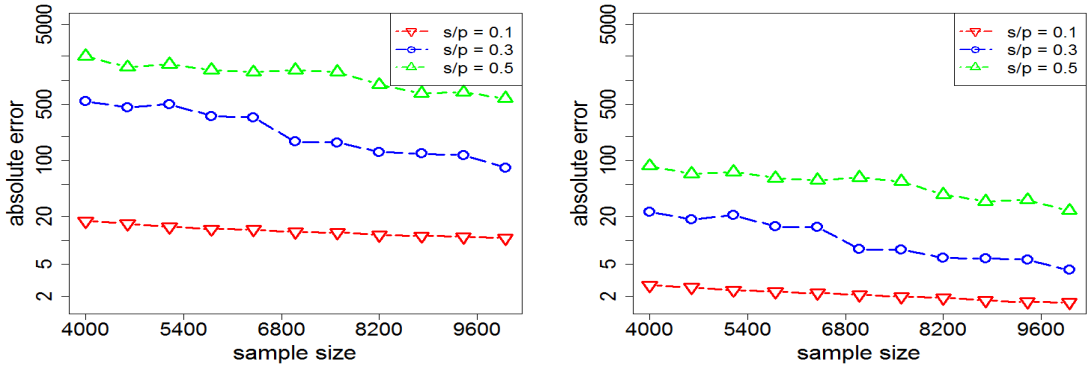

Fig. 3.3. Estimation error for different sample sizes. The left panel is for initial estimation error and the right panel is for final estimation error.

Next, we demonstrate that the low-rank tensor parameter $\mathscr{T}^*$ with randomly generated factors $\boldsymbol{\beta}_k^*$ satisfies the incoherence Condition 3 with high probability. Set the CP-rank $K = 3$ and the sparsity level $s/p = 0.3$ with the dimension p ranging from 10 to 2000. We compute the incoherence parameter Γ defined in Condition 3. The left panel of Figure 3.4 shows that the incoherence parameter Γ decays in a polynomial rate as s grows, which matches the bound in Condition 3. Recall we also provide theoretical justification on this point in Lemma 28.

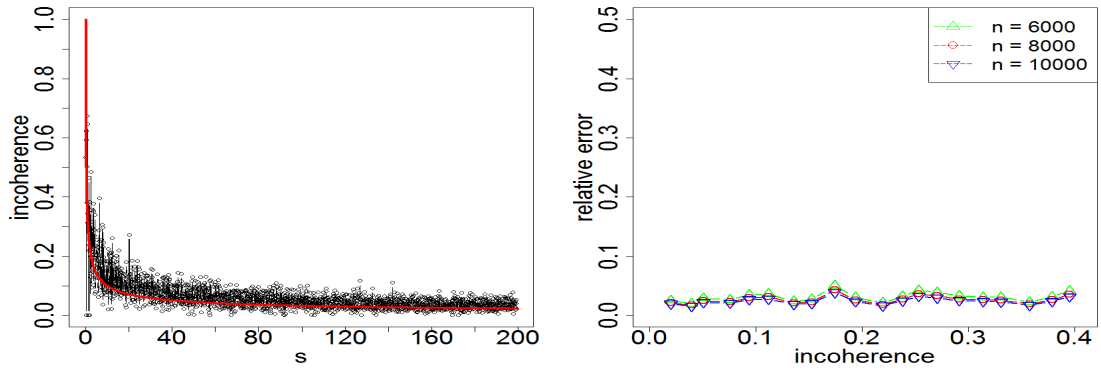

Fig. 3.4. Left panel: incoherence parameter Γ with varying sparsity. Here, the red line corresponds to the rate $\sqrt{s}$ required in the theoretical analysis. Right panel: average relative estimation error for tensors with varying incoherence.

We further examine the performance of the proposed algorithm when the incoherence condition required in the theoretical analysis fails to hold. Specifically, we set the CP-rank $K = 3$, $p = 30$, and the sparsity level $s/p = 0.3$. We construct enormous copies of tensor parameter $\mathscr{T}_j^*$ with i.i.d. standard normal factor vectors $\boldsymbol{\beta}_k^*$. For each $\mathscr{T}_j^*$, we calculate the incoherence Γ_j defined in Condition 3, and then manually pick 40 parameter tensors $\mathscr{T}_{j'}^*$ such that

$$0.01 \cdot (j' - 1) \leq \Gamma_{j'} \leq 0.01 \cdot j' \quad \text{for} \quad j' = \{1, 2, \ldots, 40\}.$$

By this construction, we obtain a set of tensor parameter $\{\mathscr{T}_{j'}^*\}$ with incoherence uniformly varying from 0 to 0.4. The right panel of Figure 3.4 plots the relative error for estimating $\mathscr{T}^*$ based on observations from cubic sketchings of $\mathscr{T}_{j'}^*$ based on 1000 repetitions. We can see that the proposed algorithm achieves small relative errors even when the true factors are highly coherent.

Moreover, we consider another setting with Laplace distributed noise which is a sub-exponential random variable. Suppose $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} Lap(\sigma)$ with density $f(x) = \frac{1}{\sigma}\exp(-2|x|/\sigma)$. With $n = 3000$, $p = 30$, and varying values of σ, the average estimation error and its comparison with Gaussian noise setting are provided in Figure

3.5. We note that the estimation errors under Laplace noise are slightly higher than those under Gaussian noise.

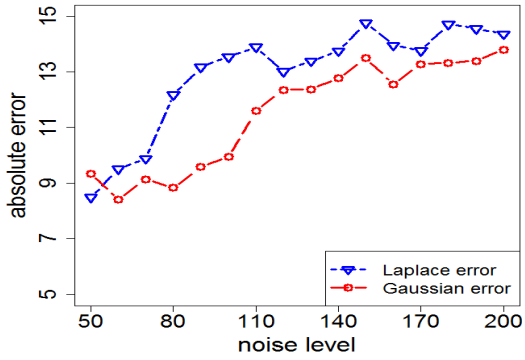

Fig. 3.5. Comparison of estimation errors between Laplace error and Gaussian error.

Next, we compare the estimation errors of initial and final estimators for different ranks and sample sizes. First we set $K = 3, p = 30, s/p = 0.3$ and consider the noiseless setting. It is clear from Figure 3.6 that the initialization error decays sufficiently, but does not converge to zero as sample size n grows. This result matches our theoretical findings in Theorem 4. As discussed in Remark 3, the initial stage may yield an inconsistent estimator due to the incoherence among $\boldsymbol{\beta}_k$'s. After sufficient steps of thresholded gradient descent (Steps 3 and 4 in Algorithm 2), the initial estimator is refined to lead to the final estimate that is proven to be minimax-optimal. Thus, we evaluate and compare estimation errors for both initial and final estimators for $K = 3$ or 5 and growing sample sizes n. We can see from the right panel of Figure 3.6, the final estimator is more stable and accurate compared with the initial one, which illustrates the merit of thresholded gradient descent step of the proposed procedure.

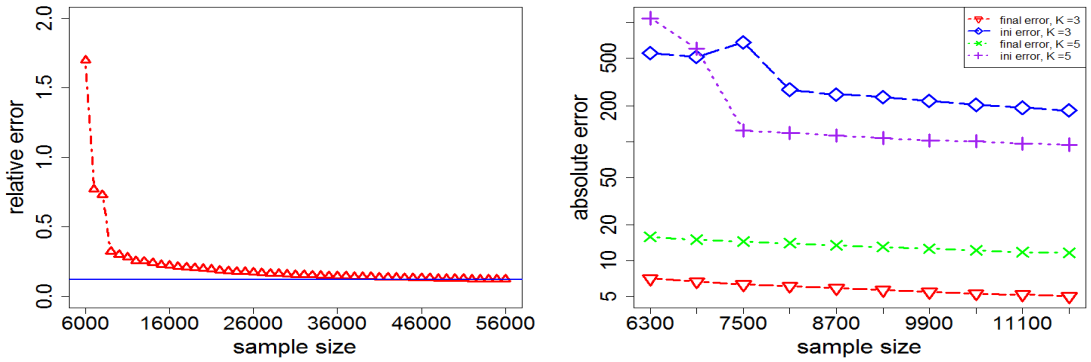

Fig. 3.6. Log absolute estimation error of initial estimation error (left panel) and initialization/final estimation error comparisons (right panel).

Last but not the least, we compare the performance of our method with the alternating least square (ALS)-based tensor regression method (32). We specifically consider two schemes for the initialization of ALS: (a) $\{\boldsymbol{\beta}_k^{(0)}\}$ were generated as i.i.d. standard Gaussian (cold start), and (b) $\{\boldsymbol{\beta}_k^{(0)}\}$ were generated from the proposed Algorithm 1 (warm start). Setting $K = 2$, $s/p = 0.2$, $p = 30$, $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, 200^2)$, we applied both our proposed procedure and the ALS-based algorithm, and recorded average estimation errors with standard deviations for both initial and final estimators in Table 3.6. From the result, one can see our proposed algorithm significantly outperforms the ALS proposed by (32) under both cold and warm start schemes. The main reason was pointed out in Remark 5: the cubic sketchings setting possesses distinct aspects compared with the i.i.d. random Gaussian sketching setting, so that the method proposed by (32) does not exactly fit.

3.7 Discussions

The current paper focuses on the third order tensor estimation. But, all the results can be extended to higher-order case via high-order sketchings as follows. To be specific, suppose

$$y_i = \langle \mathscr{T}^*, \boldsymbol{x}_i^{\otimes m} \rangle + \epsilon_i, \quad i = 1, \ldots, n,$$

Table 3.1.
The estimation error and the standard deviation (in subscript) of the proposed method and ALS-based method.

Sample size	ours	warm start	cold start	initial
$n = 4000$	$4.023_{0.135}$	$32.828_{1.798}$	$37.785_{1.233}$	$38.032_{1.748}$
$n = 5000$	$1.945_{0.097}$	$32.346_{2.343}$	$36.962_{2.106}$	$33.716_{1.786}$
$n = 6000$	$1.773_{0.092}$	$22.220_{1.215}$	$59.972_{3.407}$	$25.579_{1.483}$

where $\mathscr{T}^* \in (\mathbb{R}^p)^{\otimes m}$ is an order-m, sparse, and low-rank tensor. In order to estimate $\mathscr{T}^*$ from $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, one can first generalize Theorem 9 to construct the order-m tensor moment estimate for the initial stage, by noting that the score function $\mathcal{S}_m(\boldsymbol{x})$ and the density function $p(\boldsymbol{x})$ satisfy a nice recursive equation:

$$\mathcal{S}_m(\boldsymbol{x}) := -\mathcal{S}_{m-1}(\boldsymbol{x}) \circ \nabla \log p(\boldsymbol{x}) - \nabla \mathcal{S}_{m-1}(\boldsymbol{x}).$$

Then, one can similarly perform high-order sparse tensor decomposition and thresholded gradient descent to estimate $\mathscr{T}^*$. On the theoretical side, by a careful generalization of the truncation argument and $\psi_{(2/m)}$-norm concentration inequality, we can similarly show under mild conditions, when $n \geq C(\log n)^m (s \log p)^{m/2}$, the proposed procedure achieves the following rate of convergence with high probability,

$$\left\| \widetilde{\mathscr{T}} - \mathscr{T}^* \right\|_F^2 \lesssim \sigma^2 \frac{Kms \log(p/s)}{n},$$

where $C > 0$ is some constant which does not depend on n, p, K, m and σ^2. The minimax optimality can be shown similarly.

3.8 Proofs

In this section, we provide detailed proofs for empirical moment estimator and concentration results in Sections 3.8.1 and 3.8.2.

3.8.1 Moment Calculation

We first introduce three lemmas to show that the empirical moment based tensors (3.2.6), (3.4.5), and (3.5.4) are all unbiased estimators for the target low-rank tensor in the corresponding scenarios. Detail proofs of three lemmas are postponed to Sections 3.9.5, 3.9.6 and 3.9.7 in the supplementary materials.

Lemma 3 (Unbiasedness of moment estimator under non-symmetric sketchings). For non-symmetric tensor estimation model (3.5.1) & (3.5.2), define the empirical-moment-based tensor $\mathcal{T}$ by

$$\mathcal{T} := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i.$$

Then $\mathcal{T}$ is an unbiased estimator for $\mathscr{T}^*$, i.e.,

$$\mathbb{E}(\mathcal{T}) = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*.$$

The extension to the symmetric case is non-trivial due to the dependency among three identical sketching vectors. We borrow the idea of high-order Stein's identity, which was originally proposed in (90). To fix the idea, we present only third order result for simplicity. The extension to higher-order is straightforward.

Theorem 9 (Third-order Stein's Identity, (90)). Let $\boldsymbol{x} \in \mathbb{R}^p$ be a random vector with joint density function $p(\boldsymbol{x})$. Define the third order score function $\mathcal{S}_3(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}^{p\times p\times p}$ as $\mathcal{S}_3(\boldsymbol{x}) = -\nabla^3 p(\boldsymbol{x})/p(\boldsymbol{x})$. Then for continuously differentiable function $G(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}$, we have

$$\mathbb{E}\left[G(\boldsymbol{x}) \cdot \mathcal{S}_3(\boldsymbol{x})\right] = \mathbb{E}\left[\nabla^3 G(\boldsymbol{x})\right]. \tag{3.8.1}$$

In general, the order-m high-order score function is defined as

$$\mathcal{S}_m(\boldsymbol{x}) = (-1)^m \frac{\nabla^m p(\boldsymbol{x})}{p(\boldsymbol{x})}.$$

Interestingly, the high-order score function has a recursive differential representation

$$\mathcal{S}_m(\boldsymbol{x}) := -\mathcal{S}_{m-1}(\boldsymbol{x}) \circ \nabla \log p(\boldsymbol{x}) - \nabla \mathcal{S}_{m-1}(\boldsymbol{x}), \tag{3.8.2}$$

with $\mathcal{S}_0(\boldsymbol{x}) = 1$. This recursive form is helpful for constructing unbiased tensor estimator under symmetric cubic sketchings. Note that the first order score function $\mathcal{S}_1(\boldsymbol{x}) = -\nabla \log p(\boldsymbol{x})$ is the same as score function in Lemma 26 (Stein's lemma (96)). The proof of Theorem 9 relies on iteratively applying the recursion representation of score function (3.8.2) and the first-order Stein's lemma (Lemma 26). We provide the detailed proof in Section 3.9.4 for the sake of completeness.

In particular, if $\boldsymbol{x}$ follows a standard Gaussian vector, each order score function can be calculated based on (3.8.2) as follows,

$$\begin{aligned} &\mathcal{S}_1(\boldsymbol{x}) = \boldsymbol{x}, \mathcal{S}_2(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{x} - I_{d\times d}, \\ &\mathcal{S}_3(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x} - \sum_{j=1}^{p} \Big(\boldsymbol{x} \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{x} \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{x} \Big). \end{aligned} \tag{3.8.3}$$

Interestingly, if we let $G(\boldsymbol{x}) = \sum_{k=1}^{K} \eta_k^* (\boldsymbol{x}^\top \boldsymbol{\beta}_k^*)^3$, then

$$\frac{1}{6}\nabla^3 G(\boldsymbol{x}) = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*, \tag{3.8.4}$$

which is exactly $\mathscr{T}^*$. Connecting this fact with (3.8.1), we are able to construct the unbiased estimator in the following lemma through high-order Stein's identity.

Lemma 4 (Unbiasedness of moment estimator under symmetric sketchings). Consider the symmetric tensor estimation model (3.2.1) & (3.3.9). Define the empirical first-order moment $\boldsymbol{m}_1 := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i$. If we further define an empirical third-order-moment-based tensor $\mathcal{T}_s$ by

$$\mathcal{T}_s := \frac{1}{6}\Big[\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i - \sum_{j=1}^{p} \Big(\boldsymbol{m}_1 \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{m}_1 \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{m}_1 \Big)\Big],$$

then

$$\mathbb{E}(\mathcal{T}_s) = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*.$$

Proof. Note that $y_i = G(\boldsymbol{x}_i) + \epsilon_i$. Then we have

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} y_i \mathcal{S}_3(\boldsymbol{x})\Big) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} (G(\boldsymbol{x}_i) + \epsilon_i)\mathcal{S}_3(\boldsymbol{x}_i)\Big),$$

where $\mathcal{S}_3(\boldsymbol{x})$ is defined in (3.8.3). By using the conclusion in Theorem 9 and the fact (3.8.4), we obtain

$$\mathbb{E}(\mathcal{T}_s) = \mathbb{E}\Big(\frac{1}{6n}\sum_{i=1}^{n} y_i \mathcal{S}_3(\boldsymbol{x})\Big) = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*,$$

since ϵ_i is independent of $\boldsymbol{x}_i$. This ends the proof. ■

Although the interaction effect model (3.4.1) is still based on symmetric sketchings, we need much more careful construction for the moment-based estimator, since the first coordinate of the sketching vector is always constant 1. We give such an estimator in the following lemma.

Lemma 5 (Unbiasedness of moment estimator in interaction model). For interaction effect model (3.4.1), construct the empirical moment based tensor $\mathcal{T}_{s'}$ as following

- For $i, j, k \neq 0$, $\mathcal{T}_{s'[i,j,k]} = \mathcal{T}_{s[i,j,k]}$. And $\mathcal{T}_{s'[i,j,0]} = \mathcal{T}_{s[i,j,0]}, \mathcal{T}_{s'[0,j,k]} = \mathcal{T}_{s[0,j,k]}, \mathcal{T}_{s'[i,0,k]} = \mathcal{T}_{s[i,0,k]}$.
- For $i \neq 0$, $\mathcal{T}_{s'[0,0,i]} = \mathcal{T}_{s'[0,i,0]} = \mathcal{T}_{s'[i,0,0]} = \frac{1}{3}\mathcal{T}_{s[0,0,i]} - \frac{1}{6}(\sum_{k=1}^{p} \mathcal{T}_{s[k,k,i]} - (p+2)a_i)$.
- $\mathcal{T}_{s'[0,0,0]} = \frac{1}{2p-2}(\sum_{k=1}^{p} \mathcal{T}_{s[0,k,k]} - (p+2)\mathcal{T}_{s[0,0,0]})$.

The $\mathcal{T}_{s'}$ is an unbiased estimator for $\mathcal{B}$, i.e.,

$$\mathbb{E}(\mathcal{T}_{s'}) = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k.$$

3.8.2 Proofs of Lemmas 1 and 2: Concentration Inequalities

We aim to prove Lemmas 1 and 2 in this subsection. These two lemmas provide key concentration inequalities of the theoretical analysis for the main result. Before going into technical details, we introduce a quasi-norm called ψ_α-norm.

Definition 2 (ψ_α-norm (57)). The ψ_α-norm of any random variable X and $\alpha > 0$ is defined as

$$\|X\|_{\psi_\alpha} := \inf\Big\{C \in (0,\infty) : \ \mathbb{E}[\exp(|X|/C)^\alpha] \leq 2\Big\}.$$

Particularly, a random variable who has a bounded ψ_2-norm or bounded ψ_1-norm is called sub-Gaussian or sub-exponential random variable, respectively. Next lemma provides an upper bound for the p-th moment of sum of random variables with bounded ψ_α-norm.

Lemma 6 . Suppose $X_1, \ldots, X_n$ are n independent random variables satisfying $\|X_i\|_{\psi_\alpha} \leq b$ with $\alpha > 0$, then for all $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $p \geq 2$,

$$\begin{aligned} &\Big(\mathbb{E}\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}(\sum_{i=1}^n a_i X_i)\Big|^p\Big)^{\frac{1}{p}} \\ \leq &\begin{cases} C_1(\alpha) b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty\big), & \text{if } 0 < \alpha < 1; \\ C_2(\alpha) b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\big), & \text{if } \alpha \geq 1. \end{cases} \end{aligned} \tag{3.8.5}$$

where $1/\alpha^* + 1/\alpha = 1$, $C_1(\alpha), C_2(\alpha)$ are some absolute constants only depending on α.

If $0 < \alpha < 1$, (3.8.5) is a combination of Theorem 6.2 in (97) and the fact that the p-th moment of a Weibull variable with parameter α is of order $p^{1/\alpha}$. If $\alpha \geq 1$, (3.8.5) follows from a combination of Corollaries 2.9 and 2.10 in (98). Continuing with standard symmetrization arguments, we reach the conclusion for general random variables. When $\alpha = 1$ or 2, (3.8.5) coincides with standard moment bounds for a sum of sub-Gaussian and sub-exponential random variables in (82). The detailed proof of Lemma 6 is postponed to Section 3.9.8.

When $0 < \alpha < 1$, by Chebyshev's inequality, one can obtain the following exponential tail bound for the sum of random variables with bounded ψ_α-norm. This lemma generalizes the Hoeffding-type concentration inequality for sub-Gaussian random variables (see, e.g. Proposition 5.10 in (82)), and Bernstein-type concentration inequality for sub-exponential random variables (see, e.g. Proposition 5.16 in (82)).

Lemma 7 . Suppose $0 < \alpha < 1$, $X_1, \ldots, X_n$ are independent random variables satisfying $\|X_i\|_{\psi_\alpha} \leq b$. Then there exists absolute constant $C(\alpha)$ only depending on α such that for any $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $0 < \delta < 1/e^2$,

$$\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}(\sum_{i=1}^n a_i X_i)\Big| \leq C(\alpha) b\|\boldsymbol{a}\|_2 (\log \delta^{-1})^{1/2} + C(\alpha) b\|\boldsymbol{a}\|_\infty (\log \delta^{-1})^{1/\alpha}$$

with probability at least $1-\delta$.

Proof. For any $t>0$, by Markov's inequality,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n a_iX_i-\mathbb{E}\Big(\sum_{i=1}^n a_iX_i\Big)\Big|\geq t\Big)=\mathbb{P}\Big(\Big|\sum_{i=1}^n a_iX_i-\mathbb{E}\Big(\sum_{i=1}^n a_iX_i\Big)\Big|^p\geq t^p\Big)$$
$$\leq\frac{\mathbb{E}\Big|\sum_{i=1}^n a_iX_i-\mathbb{E}\Big(\sum_{i=1}^n a_iX_i\Big)\Big|^p}{t^p}\leq\frac{C(\alpha)^pb^p\Big(\sqrt{p}\|\boldsymbol{a}\|_2+p^{1/\alpha}\|\boldsymbol{a}\|_\infty\Big)^p}{t^p},$$

where the last inequality is from Lemma 6. We set t such that $\exp(-p)=C(\alpha)^pb^p(\sqrt{p}\|\boldsymbol{a}\|_2+p^{1/\alpha}\|\boldsymbol{a}\|_\infty)^p/t^p$. Then for $p\geq 2$,

$$\Big|\sum_{i=1}^n a_iX_i-\mathbb{E}\Big(\sum_{i=1}^n a_iX_i\Big)\Big|\leq eC(\alpha)b\Big(\sqrt{p}\|\boldsymbol{a}\|_2+p^{1/\alpha}\|\boldsymbol{a}\|_\infty\Big)$$

holds with probability at least $1-\exp(-p)$. Letting $\delta=\exp(-p)$, we have that for any $0<\delta<1/e^2$,

$$\Big|\sum_{i=1}^n a_iX_i-\mathbb{E}\Big(\sum_{i=1}^n a_iX_i\Big)\Big|\leq C(\alpha)b\Big(\|\boldsymbol{a}\|_2(\log\delta^{-1})^{1/2}+\|\boldsymbol{a}\|_\infty(\log\delta^{-1})^{1/\alpha}\Big),$$

holds with probability at least $1-\delta$. This ends the proof. ■

The next lemma provides an upper bound for the product of random variables in ψ_α-norm.

Lemma 8 (ψ_α for product of random variables). Suppose $X_1,\ldots,X_m$ are m random variables (not necessarily independent) with ψ_α-norm bounded by $\|X_j\|_{\psi_\alpha}\leq K_j$. Then the $\psi_{\alpha/m}$-norm of $\prod_{j=1}^m X_j$ is bounded as

$$\Big\|\prod_{j=1}^m X_j\Big\|_{\psi_{\alpha/m}}\leq\prod_{j=1}^m K_j.$$

Proof. For any $\{x_j\}_{j=1}^m$ and $\alpha>0$, by using the inequality of arithmetic and geometric means we have

$$\Big(|\prod_{j=1}^m\frac{x_j}{K_j}|\Big)^{\alpha/m}=\Big(\prod_{j=1}^m|\frac{x_j}{K_j}|^\alpha\Big)^{1/m}\leq\frac{1}{m}\sum_{j=1}^m|\frac{x_j}{K_j}|^\alpha.$$

Since exponential function is a monotone increasing function, it shows that

$$\begin{aligned}\exp\Big(|\prod_{j=1}^{m}\frac{x_j}{K_j}|\Big)^{\alpha/m} &\leq \exp\Big(\frac{1}{m}\sum_{j=1}^{m}|\frac{x_j}{K_j}|^{\alpha}\Big)\\ =\Big(\prod_{j=1}^{m}\exp(|\frac{x_j}{K_j}|^{\alpha})\Big)^{1/m} &\leq \frac{1}{m}\sum_{j=1}^{m}\exp\Big(|\frac{x_j}{K_j}|^{\alpha}\Big).\end{aligned} \tag{3.8.6}$$

From the definition of ψ_α-norm, for $j = 1, 2, \ldots, m$, each individual X_j has

$$\mathbb{E}\Big(\exp(\frac{|X_j|}{K_j})^{\alpha}\Big) \leq 2. \tag{3.8.7}$$

Putting (3.8.6) and (3.8.7) together, we obtain

$$\begin{aligned}&\mathbb{E}\Big[\exp\Big(|\frac{\prod_{j=1}^{m}X_j}{\prod_{j=1}^{m}K_j}|\Big)^{\alpha/m}\Big] = \mathbb{E}\Big[\exp\Big(|\prod_{j=1}^{m}\frac{X_j}{K_j}|\Big)^{\alpha/m}\Big]\\ &\leq \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\Big[\exp\Big(|\frac{X_j}{K_j}|\Big)^{\alpha}\Big] \leq 2.\end{aligned}$$

Therefore, we conclude that the $\psi_{\alpha/m}$-norm of $\prod_{j=1}^{m}X_j$ is bounded by $\prod_{j=1}^{m}K_j$. ■

Proof of Lemma 1. Note that for any $j = 1, 2, \ldots, m$, the ψ_2-norm of $\boldsymbol{X}_j^\top\boldsymbol{\beta}_j$ is bounded by $\|\boldsymbol{\beta}_j\|_2$ (82). According to Lemma 8, the $\psi_{2/m}$-norm of $\prod_{j=1}^{m}(\boldsymbol{X}_j^\top\boldsymbol{\beta}_j)$ is bounded by $\prod_{j=1}^{m}\|\boldsymbol{\beta}_j\|_2$. Directly applying Lemma 7, we reach the conclusion. ■

Proof of Lemma 2. We first focus on the non-symmetric version and the proof follows three steps:

1. Truncate the first coordinate of $\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i}$ by a carefully chosen truncation level;

2. Utilize the high-order concentration inequality in Lemma 20 at order three;

3. Show that the bias caused by truncation is negligible.

With slightly abuse of notations, we denote a, x, y etc. as their *first coordinate* of $\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y}$ etc. Without loss of generality, we assume $p := \max\{p_1, p_2, p_3\}$. By unitary

invariance, we assume $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{e}_1$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$. Then, it is equivalent to prove

$$\begin{aligned} &\Big\|M_{\text{nsy}} - \mathbb{E}(M_{\text{nsy}})\Big\|_s = \Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}x_{2i}x_{3i}\boldsymbol{x}_{1i} \circ \boldsymbol{x}_{2i} \circ \boldsymbol{x}_{3i} - \boldsymbol{e}_1 \circ \boldsymbol{e}_1 \circ \boldsymbol{e}_1\Big\|_s \\ \leq\ & C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big). \end{aligned}$$

Suppose $\boldsymbol{x}_1 \sim \mathcal{N}(0, \boldsymbol{I}_{p_1}), \boldsymbol{x}_2 \sim \mathcal{N}(0, \boldsymbol{I}_{p_2}), \boldsymbol{x}_3 \sim \mathcal{N}(0, \boldsymbol{I}_{p_3})$ and $\{\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i}\}_{i=1}^n$ are n independent samples of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$. And define a bounded event $\mathcal{G}_n$ for the first coordinate and its corresponding population version,

$$\mathcal{G}_n = \{\max_i\{|x_{1i}|, |x_{2i}|, |x_{3i}|\} \leq M\}, \mathcal{G} = \{\max\{|x_1|, |x_2|, |x_3|\} \leq M\},$$

where M is a large constant to be specified later. Decomposing $\|M_{\text{nsy}} - \mathbb{E}(M_{\text{nsy}})\|_s$ as

$$\begin{aligned} &\Big\|M_{\text{nsy}} - \mathbb{E}(M_{\text{nsy}})\Big\|_s \\ &\leq \underbrace{\Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}x_{2i}x_{3i}\boldsymbol{x}_{1i} \circ \boldsymbol{x}_{2i} \circ \boldsymbol{x}_{3i} - \mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1 \circ \boldsymbol{x}_2 \circ \boldsymbol{x}_3\big|\mathcal{G}\Big)\Big\|_s}_{M_1\text{:main term}} \\ &+ \underbrace{\Big\|\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1 \circ \boldsymbol{x}_2 \circ \boldsymbol{x}_3\big|\mathcal{G}\Big) - \boldsymbol{e}_1 \circ \boldsymbol{e}_1 \circ \boldsymbol{e}_1\Big\|_s}_{M_2\text{:bias term}}, \end{aligned}$$

we will prove that M_2 is negligible in terms of convergence rate of M_1.

Bounding M_1. For simplicity, we define $\boldsymbol{x}_1' = \boldsymbol{x}_1|\mathcal{G},\ \boldsymbol{x}_2' = \boldsymbol{x}_2|\mathcal{G},\ \boldsymbol{x}_3' = \boldsymbol{x}_3|\mathcal{G}$, and $\{\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}', \boldsymbol{x}_{3i}'\}_{i=1}^n$ are n independent samples of $\{\boldsymbol{x}_1', \boldsymbol{x}_2', \boldsymbol{x}_3'\}$. According to the law of total probability, we have

$$\begin{aligned} &\mathbb{P}\Big(M_1 \geq t\Big) \leq \mathbb{P}\Big(\mathcal{G}_n^c\Big) \\ &+ \mathbb{P}\Big(\underbrace{\Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}'\boldsymbol{x}_{1i}' \circ x_{3i}'\boldsymbol{x}_{2i}' \circ x_{i1}'\boldsymbol{x}_{3i}' - \mathbb{E}\Big(x_1'\boldsymbol{x}_1' \circ x_2'\boldsymbol{x}_2' \circ x_3'\boldsymbol{x}_3'\Big)\Big\|_s}_{M_{11}} \geq t\Big). \end{aligned}$$

According to Lemma 22, the entry of $x_{1i}'\boldsymbol{x}_{1i}', x_{2i}'\boldsymbol{x}_{2i}', x_{3i}'\boldsymbol{x}_{3i}'$ are sub-Gaussian random variable with ψ_2-norm M^2. Applying Lemma 20, we obtain

$$\mathbb{P}\Big(M_{11} \geq C_1M^6\delta_{n,s}\Big) \leq \frac{1}{p},$$

where $\delta_{n,s} = ((s\log(p/s))^3/n^2)^{1/2} + (s\log(p/s)/n)^{1/2}$.

On the other hand,

$$\mathbb{P}(\mathcal{G}_n^c) \le 3\sum_{i=1}^{n}\mathbb{P}(|x_{1i}| \ge M) \le 3ne^{1-C_2M^2}$$

Putting the above bounds together, we obtain

$$\mathbb{P}\Big(M_1 \ge C_1M^6\delta_{n,s}\Big) \le 1/s + 3ne^{1-C_2M^2}.$$

By setting $M = 2\sqrt{\log n/C_2}$, the bound of M_1 reduces to

$$\mathbb{P}\Big(M_1 \ge \frac{64C_1}{C_2^3}\delta_{n,s}(\log n)^3\Big) \le \frac{1}{p} + \frac{3e}{n^3}. \tag{3.8.8}$$

Bounding M_2. There exists $\varrho \in \mathbb{S}^{p-1}$ such that

$$M_2 = \Big|\mathbb{E}\Big(x_1x_2x_3(\boldsymbol{x}_1^\top\varrho)(\boldsymbol{x}_2^\top\varrho)(\boldsymbol{x}_3^\top\varrho)\Big|\mathcal{G}\Big) - \big(\boldsymbol{e}_1^\top\varrho\big)^3\Big|.$$

Since x_{1j} is independent of x_{1k} for any $j \ne k$, $\mathbb{E}(x_1(\boldsymbol{x}_1^\top\varrho)|\mathcal{G}) = \mathbb{E}(x_1^2\varrho_1|\mathcal{G})$. Then

$$\begin{aligned} M_2 &= \Big|\mathbb{E}\Big(x_1^2x_2^2x_3^2\varrho_1^3\Big|\mathcal{G}\Big) - \varrho_1^3\Big| \\ &= \Big|\varrho_1^3\mathbb{E}\Big(x_1^2\Big||x_1| \le M\Big)\mathbb{E}\Big(x_2^2\Big||x_2| \le M\Big)\mathbb{E}\Big(x_3^2\Big||x_3| \le M\Big) - \varrho_1^3\Big|, \end{aligned}$$

where the second equation comes from the independence among each coordinate of $\{\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i}\}$.

By the basic property of Gaussian random variable, we can show

$$1 \ge \mathbb{E}\Big(x_i^2\Big||x_i| \le M\Big) \ge 1 - 2Me^{-M^2/2}, \quad i = 1, 2, 3.$$

Plugging them into M_2, we have

$$\begin{aligned} M_2 &\le \Big|\varrho_1^3\Big|\Big|\Big(1 - 2Me^{-M^2/2}\Big)^3 - 1\Big| \\ &\le \Big|12M^2e^{-M^2} - 6Me^{-M^2/2} - 8M^3e^{-3M^2/2}\Big| \\ &\le \Big|26M^3e^{-M^2/2}\Big|, \end{aligned}$$

where the second inequality is due to $\|\varrho\|_2^2 = 1$ and the last inequality holds for a large $M > 0$. By the choice of $M = 2\sqrt{\log n/C_2}$, we have $M_2 \le 208/C_2^{3/2}(\log n)^{\frac{3}{2}}/n^2$ for some constant C_2. When n is large, this rate is negligible comparing with (3.8.8)

Bounding M: We put the upper bounds of M_1 and M_2 together. After some adjustments for absolute constant, it suffices to obtain

$$\mathbb{P}\Big(M_1 + M_2 \leq C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big)\Big) \geq 1 - \frac{10}{n^3}.$$

This concludes the proof of non-symmetric part. The proof of symmetric part remains similar and thus is omitted here. ■

3.9 Additional Results

3.9.1 Proof of Theorem 4: Initialization Effect

Theorem 4 gives an approximation error upper bound for the sparse-tensor-decomposition-based initial estimator. In Step I of Section 3.2.1, the original problem can be reformatted to a version of tensor denoising:

$$\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}, \quad \text{where} \quad \mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s). \tag{3.9.1}$$

The key difference between our model (3.9.1) and recent work is that $\mathcal{E}$ arises from empirical moment approximation, rather than the random observation noise considered in (91) and (51). Next lemma gives an upper bound for the approximation error.

Lemma 9 (Approximation error of $\mathcal{T}_s$). Recall that $\mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s)$, where $\mathcal{T}_s$ is defined in (3.2.6). Suppose Condition 4 is satisfied and $s \leq d \leq Cs$. Then

$$\|\mathcal{E}\|_{s+d} \;\leq\; 2C_1\sum_{k=1}^{K}\eta_k^*\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big)(\log n)^4 \tag{3.9.2}$$

with probability at least $1 - 5/n$ for some uniform constant C_1.

Next we denote the following quantity for simplicity,

$$\gamma = C_2\min\Big\{\frac{R^{-1}}{6} - \frac{\sqrt{K}}{s}, \frac{R^{-1}}{4\sqrt{5}} - \frac{2}{\sqrt{s}}\Big(1 + \sqrt{\frac{K}{s}}\Big)^2\Big\}, \tag{3.9.3}$$

where R is the singular value ratio, K is the CP-rank, s is the sparsity parameter, Γ is the incoherence parameter and C_2 is uniform constant.

Next lemma provides theoretical guarantees for sparse tensor decomposition method.

Lemma 10 . Suppose that the symmetric tensor denoising model (3.9.1) satisfies Conditions 1, 2 and 3 (i.e., the identifiability, parameter space and incoherence). Assume the number of initializations $L \geq K^{C_3\gamma^{-4}}$ and the number of iterations $N \geq C_4 \log\left(\gamma / \left(\frac{1}{\eta^*_{\min}}\|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2\right)\right)$ for constants C_3, C_4, the truncation parameter $s \leq d \leq Cs$. Then the sparse-tensor-decomposition-based initialization satisfies

$$\max\left\{\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \leq \frac{C_4}{\eta^*_{\min}}\|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2, \tag{3.9.4}$$

for any $k \in [K]$.

The proof of Lemma 10 essentially follows Theorem 3.9 in (51), we thus omit the detailed proof here. The upper bound in (3.9.4) contains two terms: $\frac{C_4}{\eta^*_{\min}}\|\mathcal{E}\|_{s+d}$ and $\sqrt{K}\Gamma^2$, which are due to the empirical moment approximation and the incoherence among different $\boldsymbol{\beta}_k$, respectively.

Although the sparse tensor decomposition is not optimal in statistical rate, it does offer a reasonable initial estimation provided enough samples. Equipped with (3.9.2) and Condition 2, the right side of (3.9.4) reduces to

$$\begin{aligned} &\frac{C_4}{\eta^*_{\min}}\|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2 \\ \leq\ & 2C_1C_4KR\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big)(\log n)^4 + \sqrt{K}\Gamma^2, \end{aligned}$$

with probability at least $1 - 5/n$. Denote $C_0 = 4 \cdot 2160 \cdot C_1C_4$. Using Conditions 3 and 5, we reach the conclusion that

$$\max\left\{\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \leq K^{-1}R^{-2}/2160,$$

with probability at least $1 - 5/n$. ■

3.9.2 Proof of Theorem 3: Gradient Update

We first introduce the following lemma to illustrate the improvement of one step thresholded gradient update under suitable conditions. The error bound includes two parts: the optimization error that describes one step effect for gradient update, and the statistical error that reflects the random noise effect. The proof of Lemma 11 is given in Section 3.9.10 in the supplementary materials. For notation simplicity, we drop the superscript of $\eta_k^{(0)}$ in the following proof.

Lemma 11 . Let $t \geq 0$ be an integer. Suppose Conditions 1-5 hold and $\{\boldsymbol{\beta}_k^{(t)}, \eta_k\}$ satisfies the following upper bound

$$\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \leq 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2, \ \max_{k\in[K]} \left|\eta_k - \eta_k^*\right| \leq \varepsilon_0, \tag{3.9.5}$$

with probability at least $1 - \mathcal{O}(K/n)$, where $\varepsilon_0 = K^{-1}R^{-\frac{4}{3}}/2160$. As long as the step size μ satisfies

$$0 < \mu \leq \mu_0 = \frac{32R^{-20/3}}{3K[220 + 270K]^2}, \tag{3.9.6}$$

then $\{\boldsymbol{\beta}_k^{(t+1)}\}$ can be upper bounded as

$$\begin{aligned}
&\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \\
&\leq \underbrace{\left(1 - 32\mu K^{-2}R^{-\frac{8}{3}}\right) \sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2}_{\text{optimization error}} + \underbrace{2C_0\mu^2K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2 s\log p}{n}}_{\text{statistical error}},
\end{aligned}$$

with probability at least $1 - \mathcal{O}(Ks/n)$.

In order to apply Lemma 11, we prove that the required condition (3.9.5) holds at every iteration step t by induction. When $t = 0$, by (3.3.2) and Condition 2,

$$\left\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\right\|_2 \leq \varepsilon_0, \ \left|\eta_k - \eta_k^*\right| \leq \varepsilon_0, \ \text{for } k \in [K],$$

holds with probability at least $1-\mathcal{O}(1/n)$. Since the initial estimator output by first stage is normalized, i.e., $\|\boldsymbol{\beta}_k^{(0)}\|_2 = \|\boldsymbol{\beta}_k^*\|_2 = 1$, by triangle inequality we have

$$\begin{aligned}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2 &\leq \left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^{(0)} + \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2 \\ &\leq |\sqrt[3]{\eta_k} - \sqrt[3]{\eta_k^*}| + \sqrt[3]{\eta_k^*}\left\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\right\|_2.\end{aligned}$$

Note that

$$\left|\sqrt[3]{\eta_k} - \sqrt[3]{\eta_k^*}\right| \leq \frac{\varepsilon_0}{(\sqrt[3]{\eta_k})^2 + \sqrt[3]{\eta_k\eta_k^*} + (\sqrt[3]{\eta_k^*})^2} \leq \varepsilon_0\sqrt[3]{\eta_k^*}.$$

This implies

$$\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2 \leq 2\sqrt[3]{\eta_k^*}\varepsilon_0,$$

with probability at least $1-\mathcal{O}(1/n)$. Taking the summation over $k\in[K]$, we have

$$\sum_{k=1}^K \left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \leq \sum_{k=1}^K 4\eta_k^{*\frac{2}{3}}\varepsilon_0^2 \leq 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2,$$

with probability at least $1-\mathcal{O}(K/n)$, which means (3.9.5) holds for $t=0$.

Suppose (3.9.5) holds at the iteration step $t-1$, which implies

$$\begin{aligned}&\sum_{k=1}^K \left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \\ \leq&\left(1-32\mu K^{-2}R^{-\frac{8}{3}}\right)\sum_{k=1}^K \left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t-1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 + \mu 2C_0K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*\frac{4}{3}}\frac{\sigma^2 s\log p}{n} \\ \leq&4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2 - \mu\left(128KR^{-\frac{8}{3}}\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2 - 2C_0K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*\frac{4}{3}}\frac{\sigma^2 s\log p}{n}\right).\end{aligned}$$

Since Condition 5 automatically implies

$$\frac{n}{s\log p} \geq \frac{C_0\sigma^2 R^{-\frac{2}{3}}\eta_{\min}^{*\frac{2}{3}}K}{64\varepsilon_0^2},$$

for a sufficiently large C_0, we can obtain

$$\sum_{k=1}^K \left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \leq 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2.$$

By induction, (3.9.5) holds at each iteration step.

Now we are able to use Lemma 11 recursively to complete the proof. Repeatedly using Lemma 11, we have for $t = 1, 2, \ldots,$

$$\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2$$

$$\leq \left(1 - 32\mu K^{-2} R^{-\frac{8}{3}}\right)^t \sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 + \frac{C_0 \eta_{\min}^{*-\frac{4}{3}}}{16} \frac{\sigma^2 s \log p}{n},$$

with probability at least $1 - \mathcal{O}(tKs/n)$. This concludes the first part of Theorem 3.

When the total number of iterations is no smaller than

$$T^* = \frac{\log(C_3 \eta_{\min}^{*-4/3} \sigma^2 s \log p) - \log(64 \eta_{\max}^{*2/3} K \varepsilon_0 n)}{\log(1 - 32\mu K^{-2} R^{-8/3})},$$

the statistical error will dominate the whole error bound in the sense that

$$\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(T^*)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \leq \frac{C_3 \eta_{\min}^{*-\frac{4}{3}}}{8} \frac{\sigma^2 s \log p}{n}, \tag{3.9.7}$$

with probability at least $1 - \mathcal{O}(T^* Ks/n)$.

The next lemma shows that the Frobenius norm distance between two tensors can be bounded by the distances between each factors in their CP decomposition. The proof of this lemma is provided in Section 3.9.11.

Lemma 12 . Suppose $\mathscr{T}$ and $\mathscr{T}^*$ have CP-decomposition $\mathscr{T} = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k$ and $\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$. If $|\eta_k - \eta_k^*| \leq c$, then

$$\left\| \mathscr{T} - \mathscr{T}^* \right\|_F^2 \leq 9(1+c) \Big(\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \Big) \Big(\sum_{k=1}^{K} (\sqrt[3]{\eta_k^*})^4 \Big)$$

Denote $\widehat{\mathscr{T}} = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)}$. Combing (3.9.7) and Lemma 12, we have

$$\begin{aligned} \left\| \widehat{\mathscr{T}} - \mathscr{T}^* \right\|_F^2 &\leq 9(1+\varepsilon_0) \frac{C_3 \eta_{\min}^{*-\frac{4}{3}}}{8} \frac{\sigma^2 s \log p}{n} K \eta_{\max}^{*\frac{4}{3}}, \\ &= \frac{9 C_3 R}{4} \frac{\sigma^2 K s \log p}{n}, \end{aligned}$$

with probability at least $1 - \mathcal{O}(TKs/n)$. By setting $C_1 = 9C_2/4$, we complete the proof of Theorem 3. ■

3.9.3 Proofs of Theorems 6 and 8: Minimax Lower Bounds

We first consider the proof for Theorem 8 on non-symmetric tensor estimation. Without loss of generality we assume $p = \max\{p_1, p_2, p_3\}$. We uniformly randomly generate $\{\Omega^{(k,m)}\}_{\substack{m=1,\dots,M\\k=1,\dots,K}}$ as MK subsets of $\{1,\dots,p\}$ with cardinality of s. Here $M > 0$ is a large integer to be specified later. Then we construct $\{\boldsymbol{\beta}^{(k,m)}\}_{\substack{m=1,\dots,M\\k=1,\dots,K}} \subseteq \mathbb{R}^p$ as

$$\boldsymbol{\beta}_j^{(k,m)} = \begin{cases} \sqrt{\lambda}, & \text{if } j \in \Omega^{(k,m)};\\ 0, & \text{if } j \notin \Omega^{(k,m)}. \end{cases}$$

$\lambda > 0$ will also be specified a little while later. Clearly, $\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \le 2s\lambda$ for any $1 \le k \le K$, $1 \le m_1, m_2 \le M$. Additionally, $|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}|$ satisfies the hyper-geometric distribution: $\mathbb{P}\left(\left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right| = t\right) = \frac{\binom{s}{t}\binom{p-s}{s-t}}{\binom{p}{s}}$.

Let $w^{(k,m_1,m_2)} = \left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|$, then for any $s/2 \le t \le s$,

$$\begin{aligned}
&\mathbb{P}\left(w^{(k,m_1,m_2)} = t\right) = \frac{\frac{s\cdots(s-t+1)}{t!} \cdot \frac{(p-s)\cdots(p-2s+t+1)}{(s-t)!}}{\frac{p\cdots(p-s+1)}{s!}} \le \binom{s}{t} \cdot \left(\frac{s}{p-s+1}\right)^t\\
\le& 2^s \left(\frac{s}{p-s+1}\right)^t \le \left(\frac{4s}{p-s+1}\right)^t.
\end{aligned}$$

Thus, if $\eta > 0$, the moment generating function of $w^{(k,m_1,m_2)} - \frac{s}{2}$ satisfies

$$\begin{aligned}
&\mathbb{E}\exp\left(\eta\left(w^{(k,m_1,m_2)} - \frac{s}{2}\right)\right)\\
\le& \exp(0)\cdot\mathbb{P}\left(w^{(k,m_1,m_2)} \le \frac{s}{2}\right) + \sum_{t=\lfloor s/2\rfloor+1}^{s} \exp\left(\eta\left(t - \frac{s}{2}\right)\right)\cdot\mathbb{P}\left(w^{(k,m_1,m_2)} = t\right)\\
\le& 1 + \sum_{t=\lfloor s/2\rfloor+1}^{s} \left(4s/(p-s+1)\right)^t \exp\left(\eta(t-s/2)\right)\\
\le& 1 + \left(4s/(p-s+1)\right)^{s/2} \frac{1}{1 - 4s/(p-s+1)\cdot e^{\eta}}.
\end{aligned}$$

By setting $\eta = \log((p-s+1)/(8s))$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} \geq \frac{3sK}{4}\right) = \mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2} \geq \frac{sK}{4}\right) \\
\leq &\frac{\mathbb{E}\exp\left(\eta(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2})\right)}{\exp\left(\eta \cdot \frac{sK}{4}\right)} = \frac{\prod_{k=1}^{K}\mathbb{E}\exp\left(\eta(w^{(k,m_1,m_2)} - \frac{s}{2})\right)}{\exp(\eta \cdot \frac{sK}{4})} \\
\leq &\left(1 + (4s/(p-s+1))^{s/2} \cdot 2\right)^{K} \exp\left(-\frac{sK}{4}\log\left(\frac{p-s+1}{8s}\right)\right) \\
\leq &\exp\left(-c_0 sK\log(p/s)\right)
\end{aligned}$$

for some small uniform constant $c_0 > 0$.

Next we choose $M = \lfloor \exp(c_0/2 \cdot sK\log(p/s)) \rfloor$. Note that

$$\begin{aligned}
&\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 = \lambda \cdot \left(\left|\Omega^{(k,m_1)} \setminus \Omega^{(k,m_2)}\right| + \left|\Omega^{(k,m_2)} \setminus \Omega^{(k,m_1)}\right|\right) \\
=&\lambda\left(\left|\Omega^{(k,m_1)}\right| + \left|\Omega^{(k,m_2)}\right| - 2\left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|\right) \\
=&2\lambda\left(s - \left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|\right),
\end{aligned}$$

then we further have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{K}\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \geq \frac{sK\lambda}{2}, \forall 1 \leq m_1 < m_2 \leq M\right) \\
=&\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} \leq \frac{3K}{4}, \forall 1 \leq m_1, < m_2 \leq M\right) \\
\geq &1 - \frac{M(M-1)}{2}\exp\left(-c_0 sK\log(p/s)\right) \\
> &1 - M^2\exp\left(-c_0 sK\log(p/s)\right) \geq 0,
\end{aligned}$$

which means there are positive probability that $\left\{\boldsymbol{\beta}^{(k,m)}\right\}_{\substack{k=1,\ldots,K \\ m=1,\ldots,M}}$ satisfy

$$\begin{aligned}
\frac{sK\lambda}{2} \leq &\min_{1\leq m_1<m_2\leq M}\sum_{k=1}^{K}\left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \\
\leq &\max_{1\leq m_1<m_2\leq M}\sum_{k=1}^{K}\left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \leq 2sK\lambda.
\end{aligned} \tag{3.9.8}$$

For the rest of the proof, we fix $\left\{\boldsymbol{\beta}^{(k,m)}\right\}_{\substack{k=1,\ldots,K \\ m=1,\ldots,M}}$ to be the set of vectors satisfying (3.9.8).

Next, recall the canonical basis $\boldsymbol{e}_k = (0, \ldots, \overbrace{1}^{k\text{-th}}, 0, \cdots, 0) \in \mathbb{R}^p$. Define

$$\mathscr{T}^{(m)} = \sum_{k=1}^{K} \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}_k \circ \boldsymbol{e}_k, \quad 1 \le m \le M.$$

For each tensor $\mathscr{T}^{(m)}$ and n i.i.d. Gaussian sketches $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i \in \mathbb{R}^p$, we denote the response

$$\boldsymbol{y}^{(m)} = \left\{ y_i^{(m)} \right\}_{i=1}^{n}, \quad y_i^{(m)} = \langle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \mathscr{T}^{(m)} \rangle + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$. Clearly, $\left(\boldsymbol{y}^{(m)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}\right)$ follows a joint distribution, which may vary based on different values of m.

In this step, we analyze the Kullback-Leibler divergence between different distribution pairs:

$$D_{KL}\left((\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}), (\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\right) := \mathbb{E}_{(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})} \frac{p(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{p(\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}.$$

Note that conditioning on fixed values of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$,

$$y_i^{(m)} \sim N\left(\sum_{k=1}^{K} (\boldsymbol{\beta}^{(k,m)\top} \boldsymbol{u}_i) \cdot (\boldsymbol{e}^{(k)\top} \boldsymbol{v}_i) \cdot (\boldsymbol{e}^{(k)\top} \boldsymbol{w}_i), \sigma^2 \right).$$

By the KL-divergence formula for Gaussian distribution,

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})} \left(\frac{p(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{p(\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})} \Big| \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \right) \\
=&\frac{1}{2} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \left(\left(\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)} \right)^\top \boldsymbol{u}_i \right) \left(\boldsymbol{e}^{(k)\top} \boldsymbol{v}_i \right) \left(\boldsymbol{e}^{(k)\top} \boldsymbol{w}_i \right) \right)^2 \sigma^{-2}.
\end{aligned}$$

Therefore, for any $m_1 \neq m_2$,

$$\begin{aligned}
&D_{KL}\left((\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}), (\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\right) \\
=&\mathbb{E}_{\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} (\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)})^\top \boldsymbol{u}_i) (\boldsymbol{e}^{(k)\top} \boldsymbol{v}_i) (\boldsymbol{e}^{(k)\top} \boldsymbol{w}_i) \right)^2 \sigma^{-2} \\
=&\frac{\sigma^{-2}}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{u}} ((\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)})^\top \boldsymbol{u}_i)^2 \mathbb{E}_{\boldsymbol{v}} (\boldsymbol{e}^{(k)\top} \boldsymbol{v}_i)^2 \mathbb{E}_{\boldsymbol{w}} (\boldsymbol{e}^{(k)\top} \boldsymbol{w}_i)^2 \\
=&\frac{n\sigma^{-2}}{2} \sum_{k=1}^{K} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \le \sigma^{-2} n K s \lambda.
\end{aligned}$$

Meanwhile, for any $1 \le m_1 < m_2 \le M$,

$$\|\mathscr{T}^{(m_1)} - \mathscr{T}^{(m_2)}\|_F = \left\| \sum_{k=1}^{K} (\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}) \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)} \right\|_F$$
$$= \sqrt{\sum_{k=1}^{K} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2} \overset{(3.9.8)}{\ge} \sqrt{\frac{sK\lambda}{2}}.$$

By generalized Fano's Lemma (see, e.g., (99)),

$$\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F \ge \sqrt{\frac{sK\lambda}{2}} \left(1 - \frac{\sigma^{-2} nKs\lambda + \log 2}{\log M}\right).$$

Finally we set $\lambda = \frac{c\sigma^2}{n} \log(p/s)$ for some small constant $c > 0$, then

$$\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F^2 \ge \left(\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F\right)^2 \ge \frac{c\sigma^2 sK \log(p/s)}{n}.$$

which has finished the proof of Theorem 8.

For the proof for Theorem 6, without loss of generality we assume K is a multiple of 3. We first partition $\{1, \ldots, p\}$ into two subintervals: $I_1 = \{1, \ldots, p - K/3\}, I_2 = \{p - K/3 + 1, \ldots, p\}$, randomly generate $\{\Omega^{(k,m)}\}_{\substack{m=1,\ldots,M \\ k=1,\ldots,K/3}}$ as $(MK/3)$ subsets of $\{1, \ldots, p - K/3\}$, and construct $\{\boldsymbol{\beta}^{(k,m)}\}_{\substack{m=1,\ldots,M \\ k=1,\ldots,K}} \subseteq \mathbb{R}^{p-K/3}$ as

$$\boldsymbol{\beta}^{(k,m)} = \begin{cases} \sqrt{\lambda}, & \text{if } j \notin \Omega^{(k,m)}; \\ 0, & \text{if } j \notin \Omega^{(k,m)}. \end{cases}$$

With $M = \exp(csK \log(p/s))$ and similar techniques as previous proof, one can show there exists positive possibility that

$$\frac{sK\lambda}{6} \le \min_{1 \le m_1 < m_2 \le M} \sum_{k=1}^{K/3} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2$$
$$\le \max_{1 \le m_1 < m_2 \le M} \sum_{k=1}^{K/3} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \le \frac{2sK}{3}\lambda.$$

We then construct the following candidate symmetric tensors by blockwise design,

$$\mathcal{T}^{(m)} \in \mathbb{R}^{p \times p \times p}, \quad \begin{cases} \mathcal{T}^{(m)}_{[I_1,I_2,I_2]} = \sum_{k=1}^{K/3} \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)}, \\ \mathcal{T}^{(m)}_{[I_2,I_1,I_2]} = \sum_{k=1}^{K/3} \boldsymbol{e}^{(k)} \circ \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}^{(k)}, \\ \mathcal{T}^{(m)}_{[I_2,I_2,I_1]} = \sum_{k=1}^{K/3} \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{\beta}^{(k,m)}, \\ \mathcal{T}^{(m)}_{[I_1,I_1,I_1]}, \mathcal{T}^{(m)}_{[I_1,I_1,I_2]}, \mathcal{T}^{(m)}_{[I_1,I_2,I_1]}, \mathcal{T}^{(m)}_{[I_2,I_1,I_1]}, \mathcal{T}^{(m)}_{[I_2,I_2,I_2]} \text{ are all zeros.} \end{cases}$$

Then we can see for any $\boldsymbol{u} \in \mathbb{R}^p$,

$$\langle \mathcal{T}^{(m)}, \boldsymbol{u} \circ \boldsymbol{u} \circ \boldsymbol{u} \rangle = 3 \sum_{k=1}^{K/3} \left(\boldsymbol{\beta}^{(k,m)\top} \boldsymbol{u}_{I_1} \right) \cdot \left(\boldsymbol{e}^{(k)\top} \boldsymbol{u}_{I_2} \right)^2 .$$

The rest of the proof essentially follows from the proof of Theorem 8. ■

3.9.4 Proof of Theorem 9: High-order Stein's Lemma

The proof of this theorem follows from the one of Theorem 6 in (90). For the sake of completeness, we restate the detail here. Applying the recursion representation of score function (3.8.2), we have

$$\begin{aligned} \mathbb{E}\Big[G(\boldsymbol{x}) \mathcal{S}_3(\boldsymbol{x}) \Big] &= \mathbb{E}\Big[G(\boldsymbol{x}) \Big(- \mathcal{S}_2(\boldsymbol{x}) \circ \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \mathcal{S}_2(\boldsymbol{x}) \Big) \Big] \\ &= -\mathbb{E}\Big[G(\boldsymbol{x}) \mathcal{S}_2(\boldsymbol{x}) \circ \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) \Big] - \mathbb{E}\Big[G(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \mathcal{S}_2(\boldsymbol{x}) \Big) \Big]. \end{aligned}$$

Then, we apply the first-order Stein's lemma (see Lemma 26) on function $G(\boldsymbol{x})\mathcal{S}_2(\boldsymbol{x})$ and obtain

$$\begin{aligned} \mathbb{E}\Big[G(\boldsymbol{x}) \mathcal{S}_3(\boldsymbol{x}) \Big] &= \mathbb{E}\Big[\nabla_{\boldsymbol{x}} \Big(G(\boldsymbol{x}) \mathcal{S}_2(\boldsymbol{x}) \Big) \Big] - \mathbb{E}\Big[G(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \mathcal{S}_2(\boldsymbol{x}) \Big) \Big] \\ &= \mathbb{E}\Big[\nabla_{\boldsymbol{x}} G(\boldsymbol{x}) \mathcal{S}_2(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \mathcal{S}_2(\boldsymbol{x}) G(\boldsymbol{x}) \Big] - \mathbb{E}\Big[G(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \mathcal{S}_2(\boldsymbol{x}) \Big) \Big] \\ &= \mathbb{E}\Big[\nabla_{\boldsymbol{x}} G(\boldsymbol{x}) \mathcal{S}_2(\boldsymbol{x}) \Big]. \end{aligned}$$

Repeating the above argument two more times, we reach the conclusion. ■

In this subsection, we present the detail proofs of moment calculation, including non-symmetric case, symmetric case, and interaction model.

3.9.5 Proof of Lemma 3

By the definition of $\{y_i\}$ in (3.5.1) & (3.5.2), we have

$$\begin{aligned}\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i\Big) &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i\Big) \\ &+ \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \eta_k^* (\boldsymbol{\beta}_{1k}^{*\top}\boldsymbol{u}_i)(\boldsymbol{\beta}_{2k}^{*\top}\boldsymbol{v}_i)(\boldsymbol{\beta}_{3k}^{*\top}\boldsymbol{w}_i)\boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i\Big).\end{aligned} \tag{3.9.9}$$

First, we observe $\mathbb{E}(\epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i) = 0$ due to the independence between ϵ_i and $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}$. Then, we consider a single component from a single observation

$$M = \mathbb{E}((\boldsymbol{\beta}_{1k}^{*\top}\boldsymbol{u}_i)(\boldsymbol{\beta}_{2k}^{*\top}\boldsymbol{v}_i)(\boldsymbol{\beta}_{3k}^{*\top}\boldsymbol{w}_i)\boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i),\ i \in [n], k \in [K].$$

For notation simplicity, we drop the subscript i for i-th observation and k for k-th component such that

$$M = \mathbb{E}\Big((\boldsymbol{\beta}_1^{*\top}\boldsymbol{u})(\boldsymbol{\beta}_2^{*\top}\boldsymbol{v})(\boldsymbol{\beta}_3^{*\top}\boldsymbol{w})\boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\Big) \in \mathbb{R}^{p_1 \times p_2 \times p_3}. \tag{3.9.10}$$

Each entry of M can be calculated as follows

$$\begin{aligned} M_{ijk} &= \mathbb{E}\Big((\boldsymbol{\beta}_1^{*\top}\boldsymbol{u})(\boldsymbol{\beta}_2^{*\top}\boldsymbol{v})(\boldsymbol{\beta}_3^{*\top}\boldsymbol{w})u_i v_j w_k\Big) \\ &= \mathbb{E}\Big((\beta_{1i}^* u_i + \sum_{m\neq i}\beta_{1m}^* u_m)u_i\Big)\mathbb{E}\Big((\beta_{2j}^* u_i + \sum_{m\neq j}\beta_{2m}^* v_m)v_j\Big) \\ &\quad \times \mathbb{E}\Big((\beta_{3k}^* w_k + \sum_{m\neq k}\beta_{3m}^* w_m)w_k\Big) \\ &= \beta_{1i}^*\beta_{2j}^*\beta_{3k}^*, \end{aligned}$$

which implies $M = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \boldsymbol{\beta}_3$. Combining with n observations and K components, we can obtain

$$\mathbb{E}(\mathcal{T}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k} \circ \boldsymbol{\beta}_{2k} \circ \boldsymbol{\beta}_{3k}.$$

This finished our proof. ■

3.9.6 Proof of Lemma 4

In this subsection, we provide an alternative and more direct proof for Lemma 4. We consider a similar single component of (3.9.10) but with a symmetric structure, namely, $M_s = \mathbb{E}\Big((\boldsymbol{\beta}^{*\top}\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big)$. Based on the symmetry of both underlying tensor and sketchings, we will verify the following three cases:

- When $i=j=k$, then

$$\begin{aligned} M_{s_{iii}} &= \mathbb{E}\Big(\beta_i^* x_i + \sum_{m\neq i}\beta_m^* x_m\Big)^3 x_i^3 \\ &= \mathbb{E}\Big(\beta_i^{*3}x_i^3 + 3\beta_i^{*2}x_i^2\big(\sum_{m\neq i}\beta_m^* x_m\big) \\ &\quad +3\beta_i^* x_i\big(\sum_{m\neq i}\beta_m^* x_m\big)^2 + \big(\sum_{m\neq i}\beta_m^* x_m\big)^3\Big)x_i^3 \\ &= 15\beta_i^{*3} + 9\beta_i^*\sum_{m\neq i}\beta_m^{*2} = 9\beta_i^* + 6\beta_i^{*3}. \end{aligned}$$

  The last equation is due to $\|\boldsymbol{\beta}^*\|_2 = 1$.

- When $i\neq j\neq k$, then

$$\begin{aligned} M_{s_{ijk}} &= \mathbb{E}\Big(\beta_i^* x_i + \beta_j^* x_j + \beta_k^* x_k + \sum_{m\neq i,j,k}\beta_m^* x_m\Big)^3 x_i x_j x_k \\ &= \mathbb{E}\big(\beta_i^* x_i + \beta_j^* x_j + \beta_k^* x_k\big)^3 x_i x_j x_k \\ &= 6\beta_i^*\beta_j^*\beta_k^*. \end{aligned}$$

- When $i=j\neq k$, then

$$\begin{aligned} M_{s_{iik}} &= \mathbb{E}\Big(\beta_i^* x_i + \beta_k^* x_k + \sum_{m\neq i,k}\beta_m^* x_m\Big)^3 x_i^2 x_k \\ &= 9\beta_i^{*2}\beta_k^* + 3\beta_k^{*3} + 3\beta_k^*\big(\sum_{m\neq i,k}\beta_m^{*2}\big) \\ &= 9\beta_i^{*2}\beta_k^* + 3\beta_k^*\big(\sum_{m\neq i}\beta_m^{*2}\big) \\ &= 3\beta_k^* + 6\beta_i^{*2}\beta_k^*. \end{aligned}$$

Therefore, it is sufficient to calculate M_s by

$$\begin{aligned} M_s =& 3\sum_{k=1}^{K}\eta_k^*\Big(\sum_{m=1}^{p}\boldsymbol{\beta}_k^*\circ\boldsymbol{e}_m\circ\boldsymbol{e}_m+\boldsymbol{e}_m\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{e}_m+\boldsymbol{e}_m\circ\boldsymbol{e}_m\circ\boldsymbol{\beta}_k^*\Big)\\ &+6\sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*. \end{aligned}$$

The first term is the bias term due to correlations among symmetric sketchings. Denote $M_1=\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{x}_i$ and note that $\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{x}_i\Big)=3\sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^*$. Therefore, the empirical first-order moment M_1 could be used to remove the bias term as follows

$$\begin{aligned} &\mathbb{E}\Big(M_s-\sum_{m=1}^{p}\Big(M_1\circ\boldsymbol{e}_m\circ\boldsymbol{e}_m+\boldsymbol{e}_m\circ M_1\circ\boldsymbol{e}_m+\boldsymbol{e}_m\circ\boldsymbol{e}_m\circ M_1\Big)\Big)\\ =\;& 6\sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*. \end{aligned}$$

This finishes our proof. ■

3.9.7 Proof of Lemma 5

As before, consider a single component first. For notation simplicity, we drop the subscript l for l-th observation and k for k-th component. Since each component is normalized, the entry-wise expectation of $(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}$ can be calculated as

$$\begin{aligned} \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{0,0,0} &= 3\beta_0-2\beta_0^3\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{0,0,i} &= 3\beta_i\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{0,i,i} &= 6\beta_0\beta_i^2+3\beta_0\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{0,i,j} &= 6\beta_0\beta_i\beta_j\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{i,i,i} &= 6\beta_i^3+9\beta_i\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{i,i,j} &= 6\beta_i^2\beta_j+3\beta_j\\ \Big[\mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{x})^3\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big]_{i,j,k} &= 6\beta_i\beta_j\beta_k. \end{aligned}$$

Due to the symmetric structure and non-randomness of first coordinate, there are bias appearing for each entry. For $i, j, k \neq 0$, we could use $\sum_{m=1}^{p}(\boldsymbol{a} \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{a} \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ \boldsymbol{a})$ to remove the bias as shown in the previous proof of Lemma 4. For the subscript involving 0, the following two calculations work for removing the bias,

$$\mathbb{E}\Big(\frac{1}{3}\mathcal{T}_s - \frac{1}{6}(\sum_{k=1}^{p}\mathcal{T}_{s,[k,k,i]} - (p+1)\boldsymbol{a}_i)\Big) = \beta_0^2\beta_i.$$
$$\mathbb{E}\Big(\frac{1}{2p-2}(\sum_{k=1}^{p}\mathcal{T}_{s[0,k,k]} - (p+2)\mathcal{T}_{s[0,0,0]})\Big) = \beta_0^3.$$

This ends the proof. ■

3.9.8 Proof of Lemma 6

Without loss of generality, we assume $\|X_i\|_{\psi_\alpha} = 1$ and $\mathbb{E}X_i = 0$ throughout this proof. Let $\beta = (\log 2)^{1/\alpha}$ and $Z_i = (|X_i| - \beta)_+$, where $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if else. For notation simplicity, we define $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ for a random variable X. The following step is to estimate the moment of linear combinations of variables $\{X_i\}_{i=1}^n$.

According to the symmetrization inequality (e.g., Proposition 6.3 of (100)), we have

$$\Big\|\sum_{i=1}^{n} a_i X_i\Big\|_p \leq 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i X_i\Big\|_p = 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i|X_i|\Big\|_p, \tag{3.9.11}$$

where $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables and we notice that $\varepsilon_i X_i$ and $\varepsilon_i|X_i|$ are identically distributed. Moreover, if $|X_i| \geq \beta$, the definition of Z_i implies that $|X_i| = Z_i + \beta$. And if $|X_i| < \beta$, we have $Z_i = 0$. Thus, we have $|X_i| \leq Z_i + \beta$ at any time and it leads to

$$2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i|X_i|\Big\|_p \quad \leq \quad 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i(\beta + Z_i)\Big\|_p. \tag{3.9.12}$$

By triangle inequality,

$$2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i(\beta + Z_i)\Big\|_p \leq 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i Z_i\Big\|_p + 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i\beta\Big\|_p. \tag{3.9.13}$$

Next, we will bound the second term of the RHS of (3.9.13). In particular, we will utilize Khinchin-Kahane inequality, whose formal statement is included in Lemma 27 for the sake of completeness. From Lemma 27 we have

$$\begin{aligned}\Big\|\sum_{i=1}^n a_i\varepsilon_i\beta\Big\|_p &\le \Big(\frac{p-1}{2-1}\Big)^{1/2}\Big\|\sum_{i=1}^n a_i\varepsilon_i\beta\Big\|_2 \\ &\le \beta\sqrt{p}\Big\|\sum_{i=1}^n a_i\varepsilon_i\Big\|_2. \end{aligned} \tag{3.9.14}$$

Since $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables, some simple calculations implies

$$\begin{aligned}\Big(\mathbb{E}\Big(\sum_{i=1}^n \varepsilon_i a_i\Big)^2\Big)^{1/2} &= \Big(\mathbb{E}\Big(\sum_{i=1}^n \varepsilon_i^2 a_i^2 + 2\sum_{1\le i<j\le n}\varepsilon_i\varepsilon_j a_i a_j\Big)\Big)^{1/2} \\ &= \Big(\sum_{i=1}^n a_i^2\mathbb{E}\varepsilon_i^2 + 2\sum_{1\le i<j\le n} a_i a_j\mathbb{E}\varepsilon_i\mathbb{E}\varepsilon_j\Big)^{1/2} \\ &= \Big(\sum_{i=1}^n a_i^2\Big)^{1/2} = \|\boldsymbol{a}\|_2. \end{aligned} \tag{3.9.15}$$

Combining inequalities (3.9.12)-(3.9.15),

$$2\Big\|\sum_{i=1}^n a_i\varepsilon_i|X_i|\Big\|_p \le 2\Big\|\sum_{i=1}^n a_i\varepsilon_i Z_i\Big\|_p + 2\beta\sqrt{p}\|\boldsymbol{a}\|_2. \tag{3.9.16}$$

Let $\{Y_i\}_{i=1}^n$ are independent symmetric random variables satisfying $\mathbb{P}(|Y_i|\ge t) = \exp(-t^\alpha)$ for all $t\ge 0$. Then we have

$$\begin{aligned}\mathbb{P}(Z_i\ge t) \le& \mathbb{P}(|X_i|\ge t+\beta) = \mathbb{P}\left(\exp(|X_i|^\alpha)\ge\exp((t+\beta)^\alpha)\right) \\ \le& (\mathbb{E}|X_i|^\alpha)\cdot\exp(-(t+\beta)^\alpha) \le 2\exp(-(t+\beta)^\alpha) \\ \le& 2\exp(-t^\alpha-\beta^\alpha) = \mathbb{P}(|Y_i|\ge t),\end{aligned}$$

which implies

$$\Big\|\sum_{i=1}^n a_i\varepsilon_i Z_i\Big\|_p \le \Big\|\sum_{i=1}^n a_i\varepsilon_i Y_i\Big\|_p = \Big\|\sum_{i=1}^n a_i Y_i\Big\|_p, \tag{3.9.17}$$

since $\varepsilon_i Y_i$ and Y_i have the same distribution due to symmetry. Combining (3.9.16) and (3.9.17) together, we reach

$$\Big\|\sum_{i=1}^n a_i X_i\Big\|_p \le 2\beta\sqrt{p}\|\boldsymbol{a}\|_2 + 2\Big\|\sum_{i=1}^n a_i Y_i\Big\|_p. \tag{3.9.18}$$

For $0 < \alpha < 1$, it follows Lemma 25 that

$$\Big\| \sum_{i=1}^{n} a_i Y_i \Big\|_p \leq C_1(\alpha)(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty), \tag{3.9.19}$$

where $C_1(\alpha)$ is some absolute constant only depending on α.

For $\alpha \geq 1$, we will combine Lemma 24 and the method of the integration by parts to pass from tail bound result to moment bound result. Recall that for every non-negative random variable X, integration by parts yields the identity

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)dt.$$

Applying this to $X = |\sum_{i=1}^n a_i Y_i|^p$ and changing the variable $t = t^p$, then we have

$$\begin{aligned} \mathbb{E}|\sum_{i=1}^{n} a_i Y_i|^p &= \int_0^\infty \mathbb{P}\Big(|\sum_{i=1}^{n} a_i Y_i| \geq t\Big) p t^{p-1} dt \\ &\leq \int_0^\infty 2\exp\Big(-c\min\Big(\frac{t^2}{\|\boldsymbol{a}\|_2^2}, \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big) p t^{p-1} dt, \end{aligned} \tag{3.9.20}$$

where the inequality is from Lemma 24 for all $p \geq 2$ and $1/\alpha + 1/\alpha^* = 1$. In this following, we bound the integral in three steps:

1. If $\frac{t^2}{\|\boldsymbol{a}\|_2^2} \leq \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (3.9.20) reduces to

$$\mathbb{E}|\sum_{i=1}^{n} a_i Y_i|^p \leq 2p\int_0^\infty \exp\Big(-c\frac{t^2}{\|\boldsymbol{a}\|_2^2}\Big)\Big) t^{p-1} dt.$$

Letting $t' = ct^2/\|\boldsymbol{a}\|_2^2$, we have

$$\begin{aligned} 2p\int_0^\infty \exp\Big(-c\frac{t^2}{\|\boldsymbol{a}\|_2^2}\Big)\Big) t^{p-1} dt &= \frac{p\|\boldsymbol{a}\|_2^p}{c^{p/2}} \int_0^\infty e^{-t'} t'^{p/2-1} dt' \\ &= \frac{p\|\boldsymbol{a}\|_2^p}{c^{p/2}}\Gamma(\frac{p}{2}) \leq \frac{p\|\boldsymbol{a}\|_2^p}{c^{p/2}}(\frac{p}{2})^{p/2}, \end{aligned}$$

where the second equation is from the density of Gamma random variable. Thus,

$$\Big(\mathbb{E}|\sum_{i=1}^{n} a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \frac{p^{1/p}}{(2c)^{1/2}}\sqrt{p}\|\boldsymbol{a}\|_2 \leq \frac{\sqrt{2}}{\sqrt{c}}\sqrt{p}\|\boldsymbol{a}\|_2. \tag{3.9.21}$$

2. If $\frac{t^2}{\|\boldsymbol{a}\|_2^2} > \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (3.9.20) reduces to

$$\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p \leq 2p\int_0^\infty \exp\Big(-c\frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big)t^{p-1}dt.$$

Letting $t' = ct^\alpha/\|\boldsymbol{a}\|_{\alpha^*}^\alpha$, we have

$$\begin{aligned} 2p\int_0^\infty \exp\Big(-c\frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big)t^{p-1}dt &= \frac{2p\|\boldsymbol{a}\|_{\alpha^*}^p}{\alpha c^{p/\alpha}}\int_0^\infty e^{-t'}t'^{p/\alpha-1}dt' \\ &= \frac{2}{\alpha}\frac{p\|\boldsymbol{a}\|_{\alpha^*}^p}{c^{p/\alpha}}\Gamma(\frac{p}{\alpha}) \leq \frac{2}{\alpha}\frac{p\|\boldsymbol{a}\|_{\alpha^*}^p}{c^{p/\alpha}}(\frac{p}{\alpha})^{p/\alpha}. \end{aligned}$$

Thus,

$$\Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \frac{2p^{1/p}}{(c\alpha)^{1/\alpha}}p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*} \leq \frac{4}{(c\alpha)^{1/\alpha}}p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}. \tag{3.9.22}$$

3. Overall, we have the following by combining (3.9.21) and (3.9.22),

$$\Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \max\Big(\sqrt{\frac{2}{c}}, \frac{4}{(c\alpha)^{1/\alpha}}\Big)\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\Big).$$

After denoting $C_2(\alpha) = \max\Big(\sqrt{\frac{2}{c}}, \frac{4}{(c\alpha)^{1/\alpha}}\Big)$, we reach

$$\Big\|\sum_{i=1}^n a_i Y_i\Big\|_p \leq C_2(\alpha)\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\Big). \tag{3.9.23}$$

Since $0 < \beta < 1$, the conclusion can be reached by combining (3.9.18),(3.9.19) and (3.9.23). ■

3.9.9 Proof of Lemma 9

Firstly, let us consider the non-symmetric perturbation error analysis. According to Lemma 3, the exact form of $\mathcal{E} = \mathcal{T} - \mathbb{E}(\mathcal{T})$ is given by

$$\mathcal{E} = \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i - \sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*.$$

We decompose it by a concentration term ($\mathcal{E}_1$) and a noise term ($\mathcal{E}_2$) as follows,

$$\mathcal{E} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^* \rangle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i - \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*}_{\mathcal{E}_1}$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i}_{\mathcal{E}_2}.$$

Bounding $\mathcal{E}_1$: For k-th componet of $\mathcal{E}_1$, we denote

$$\mathcal{E}_{1k} = \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^* \rangle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i - \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*.$$

By using Lemma 2 and $s \le d \le Cs$, it suffices to have for some absolute constant C_{11},

$$\|\mathcal{E}_{1k}\|_{s+d} \le C_{11}\delta_{n,p,s}, \text{ where } \delta_{n,p,s} = (\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big),$$

with probability at least $1 - 10/n^3$, where $\|\cdot\|_{s+d}$ is the sparse tensor spectral norm defined in (3.1.3). Equipped with the triangle inequality, the sparse tensor spectral norm for $\mathcal{E}_1$ can be bounded by

$$\|\mathcal{E}_1\|_{s+d} \le C_{11}\delta_{n,p,s}\sum_{k=1}^{K}\eta_k^*, \tag{3.9.24}$$

with probability at least $1 - 10K/n^3$.

Bounding $\mathcal{E}_2$: Note that the random noise $\{\epsilon_i\}_{i=1}^n$ is independent of sketching vector $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}$. For fixed $\{\epsilon_i\}_{i=1}^n$, applying Lemma 20, we have for some absolute constant C_{12}

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i\Big\|_{s+d} \le C_{12}\|\boldsymbol{\epsilon}\|_\infty C_{11}\delta_{n,p,s},$$

with probability at least $1 - 1/p$. According to Lemma 23, we have

$$\mathbb{P}\Big(\|\mathcal{E}_2\|_{s+d} \ge C_{12}\sigma\log n\delta_{n,p,s}\Big) \le \frac{1}{p} + \frac{3}{n} \le \frac{4}{n}. \tag{3.9.25}$$

Bounding $\mathcal{E}$: Putting (3.9.24) and (3.9.25) together, we obtain

$$\|\mathcal{E}\|_{s+d} \le \Big(C_{11}\sum_{k=1}^{K}\eta_k^* + C_{12}\sigma\log n\Big)\delta_{n,p,s},$$

with probability at least $1 - 5/n$. Under Condition 9, we have

$$\|\mathcal{E}\|_{s+d} \leq 2C_1 \sum_{k=1}^{K} \eta_k^* \delta_{n,p,s} \log n,$$

with probability at least $1 - 5/n$.

The perturbation error analysis for the symmetric tensor estimation model and the interaction effect model is similar since the empirical-first-order moment converges much faster than the empirical-third-order moment. So we omit the detailed proof here. ■

3.9.10 Proof of Lemma 11

Lemma 11 quantifies one step update for thresholded gradient update. The proof consists of two parts.

First, we evaluate an oracle estimator $\{\widetilde{\boldsymbol{\beta}}_k^{(t+1)}\}_{k=1}^K$ with known support information, which is defined as

$$\widetilde{\boldsymbol{\beta}}_k^{(t+1)} = \varphi_{\frac{\mu}{\phi} h(\boldsymbol{\beta}_k^{(t)})}\Big(\boldsymbol{\beta}_k^{(t)} - \frac{\mu}{\phi} \nabla_k \mathcal{L}(\boldsymbol{\beta}_k^{(t)})_{F^{(t)}}\Big). \tag{3.9.26}$$

Here,

- $h(\boldsymbol{\beta}_k^{(t)})$ is the k-th component of $h(\boldsymbol{B}^{(t)})$ defined in (3.2.2).
- $\nabla_{\boldsymbol{B}} \mathcal{L}(\boldsymbol{B}) = (\nabla_1 \mathcal{L}(\boldsymbol{\beta}_1), \cdots, \nabla_K \mathcal{L}(\boldsymbol{\beta}_K))$.
- $F^{(t)} = \cup_{k=1}^K F_k^{(t)}$, where $F_k^{(t)} = \mathrm{supp}(\boldsymbol{\beta}_k^*) \cup \mathrm{supp}(\boldsymbol{\beta}_k^{(t)})$.
- For a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a subset $A \subset \{1, \ldots, p\}$, we denote $\boldsymbol{x}_A \in \mathbb{R}^p$ by keeping the coordinates of $\boldsymbol{x}$ with indices in A unchanged, while changing all other components to zero.

We will show that $\widetilde{\boldsymbol{\beta}}_k^{(t+1)}$ converges as a geometric rate for optimization error and an optimal rate for statistical error. See Lemma 13 for details.

Second, we aim to prove that $\widetilde{\boldsymbol{\beta}}_k^{(t+1)}$ and $\boldsymbol{\beta}_k^{(t+1)}$ are almost equivalent with high probability. See Lemma 14 for details. For simplicity, we drop the superscript of $\boldsymbol{\beta}_k^{(t)}, F^{(t)}$ in the following proof, and denote $\widetilde{\boldsymbol{\beta}}_k^{(t+1)}$, $\boldsymbol{\beta}_k^{(t+1)}$ and $F^{(t+1)}$ by $\widetilde{\boldsymbol{\beta}}_k^{+}$, $\widetilde{\boldsymbol{\beta}}_k^{+}$ and F^{+}, respectively.

Lemma 13 . Suppose Conditions 1-5 hold. Assume (3.9.5) is satisfied and $|F| \lesssim Ks$. As long as the step size $\mu \leq 32R^{-20/3}/(3K[220+270K]^2)$, we obtain the upper bound for $\{\widetilde{\boldsymbol{\beta}}_k^{+}\}$,

$$\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \leq \Big(1-32\mu\frac{R^{-\frac{8}{3}}}{K^2}\Big)\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 + 2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n}, \tag{3.9.27}$$

with probability at least $1-(21K^2+11K+4Ks)/n$.

The proof of Lemma 13 is postponed to the Section 3.9.12. Next lemma guarantees that with high probability, $\{\boldsymbol{\beta}_k^{+}\}_{k=1}^K$ is equivalent to the oracle update $\{\widetilde{\boldsymbol{\beta}}_k^{+}\}_{k=1}^K$ with high probability.

Lemma 14 . Recall that the truncation level $h(\boldsymbol{\beta}_k)$ is defined as

$$h(\boldsymbol{\beta}_k)=\frac{\sqrt{4\log np}}{n}\sqrt{\sum_{i=1}^{n}\Big(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i\Big)^2\Big(\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2\Big)^2}. \tag{3.9.28}$$

If $|F| \lesssim Ks$, we have $\boldsymbol{\beta}_k^{+}=\widetilde{\boldsymbol{\beta}}_k^{+}$ for any $k\in[K]$ with probability at least $1-(n^2p)^{-1}$ and $F^{+}\subset F$.

The proof of Lemma 14 is postponed to the Section 3.9.12. By using Lemma 14 and induction, we have

$$F^{(t+1)}\subset\cdots F^{(1)}\subset F^{(0)}=\cup_{k=1}^{K}\mathrm{supp}(\boldsymbol{\beta}_k^*)\cup\mathrm{supp}(\boldsymbol{\beta}_k^{(0)}).$$

It implies for every t, we have $|F^{(t)}| \lesssim Ks$. Combining with Lemmas 13 and 14 together, we obtain with probability at least $1-(21K^2+11K+4Ks)/n$,

$$\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \leq \Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 + 2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n}, \tag{3.9.29}$$

This ends the proof. ■

3.9.11 Proof of Lemma 12

Based on the CP low-rank structure of true tensor parameter $\mathscr{T}^*$, we can explicitly write down the distance between $\mathscr{T}$ and $\mathscr{T}^*$ under tensor Frobenius norm as follows

$$\left\|\mathscr{T}-\mathscr{T}^*\right\|_F^2=\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K\eta_k\beta_{ki_1}\beta_{ki_2}\beta_{ki_3}-\sum_{k=1}^K\eta_k^*\beta_{ki_1}^*\beta_{ki_2}^*\beta_{ki_3}^*\Big)^2.$$

For notation simplicity, denote $\bar{\boldsymbol{\beta}}_k=\sqrt[3]{\eta_k}\boldsymbol{\beta}_k,\bar{\boldsymbol{\beta}}_k^*=\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*$. Then

$$\begin{aligned}\left\|\mathscr{T}-\mathscr{T}^*\right\|_F^2&=\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K\bar{\beta}_{ki_1}\bar{\beta}_{ki_2}\bar{\beta}_{ki_3}-\sum_{k=1}^K\bar{\beta}_{ki_1}^*\bar{\beta}_{ki_2}^*\bar{\beta}_{ki_3}^*\Big)^2\\&=\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K(\bar{\beta}_{ki_1}-\bar{\beta}_{ki_1}^*)\bar{\beta}_{ki_2}^*\bar{\beta}_{ki_3}^*+\sum_{k=1}^K\bar{\beta}_{ki_1}(\bar{\beta}_{ki_2}-\bar{\beta}_{ki_2}^*)\bar{\beta}_{ki_3}^*\\&\quad+\sum_{k=1}^K\bar{\beta}_{ki_1}\bar{\beta}_{ki_2}(\bar{\beta}_{ki_3}-\bar{\beta}_{ki_3}^*)\Big)^2=\text{RHS}.\end{aligned}$$

Since $(a+b+c)^2\le 3(a^2+b^2+c^2)$, we have

$$\begin{aligned}\text{RHS}\le&3\sum_{i_1,i_2,i_3}\Big[(\sum_{k=1}^K(\bar{\beta}_{ki_1}-\bar{\beta}_{ki_1}^*)\bar{\beta}_{ki_2}^*\bar{\beta}_{ki_3}^*)^2+(\sum_{k=1}^K\bar{\beta}_{ki_1}(\bar{\beta}_{ki_2}-\bar{\beta}_{ki_2}^*)\bar{\beta}_{ki_3}^*)^2\\&+(\sum_{k=1}^K\bar{\beta}_{ki_1}\bar{\beta}_{ki_2}(\bar{\beta}_{ki_3}-\bar{\beta}_{ki_3}^*))^2\Big].\end{aligned}$$

Equipped with Cauchy-Schwarz inequality, RHS can be further bounded by

$$\begin{aligned}\text{RHS}\quad\le\quad&3\sum_{i_1,i_2,i_3}\Big[\sum_{k=1}^K(\bar{\beta}_{ki_1}-\bar{\beta}_{ki_1}^*)^2\sum_{k=1}^K\bar{\beta}_{ki_2}^{*2}\bar{\beta}_{ki_3}^{*2}\\&+\sum_{k=1}^K(\bar{\beta}_{ki_2}-\bar{\beta}_{ki_2}^*)^2\sum_{k=1}^K\bar{\beta}_{ki_1}^2\bar{\beta}_{ki_3}^{*2}\\&+\sum_{k=1}^K(\bar{\beta}_{ki_3}-\bar{\beta}_{ki_3}^*)^2\sum_{k=1}^K\bar{\beta}_{ki_2}^2\ \bar{\beta}_{ki_1}^2\Big]\end{aligned}$$

At the same time, using $\eta_k \le (1+c)\eta_k^*$ for $k \in [K]$,

$$\begin{aligned}
\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2 \;\le\; & 3\Big[\sum_{i_1=1}^{p}\sum_{k=1}^{K}(\bar\beta_{ki_1}-\bar\beta^*_{ki_1})^2\Big(\sum_{i_2=1}^{p}\sum_{i_3=1}^{p}\sum_{k=1}^{K}\bar\beta^{*2}_{ki_2}\bar\beta^{*2}_{ki_3}\Big)\\
& +\sum_{i_2=1}^{p}\sum_{k=1}^{K}(\bar\beta_{ki_2}-\bar\beta^*_{ki_2})^2\Big(\sum_{i_1=1}^{p}\sum_{i_3=1}^{p}\sum_{k=1}^{K}\bar\beta^{2}_{ki_1}\bar\beta^{*2}_{ki_3}\Big)\\
& +\sum_{i_3=1}^{p}\sum_{k=1}^{K}(\bar\beta_{ki_3}-\bar\beta^*_{ki_3})^2\Big(\sum_{i_2=1}^{p}\sum_{i_1=1}^{p}\sum_{k=1}^{K}\bar\beta^{2}_{ki_2}\bar\beta^{2}_{ki_1}\Big)\Big]\\
= \;& 3\Big(\sum_{k=1}^{K}\|\bar{\boldsymbol\beta}_k-\bar{\boldsymbol\beta}_k^*\|_2^2\Big)\Big(\sum_{k=1}^{K}(\sqrt[3]{\eta_k^*})^4+\sum_{k=1}^{K}(\sqrt[3]{\eta_k^*})^2(\sqrt[3]{\eta_k})^2+\sum_{k=1}^{K}(\sqrt[3]{\eta_k})^4\Big)\\
\le \;& 9(1+c)\Big(\sum_{k=1}^{K}\|\bar{\boldsymbol\beta}_k-\bar{\boldsymbol\beta}_k^*\|_2^2\Big)\Big(\sum_{k=1}^{K}(\sqrt[3]{\eta_k^*})^4\Big).
\end{aligned}$$

For the non-symmetric tensor estimation model, we have

$$\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2=\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^{K}\eta_k\beta_{1ki_1}\beta_{2ki_2}\beta_{3ki_3}-\sum_{k=1}^{K}\eta_k^*\beta^*_{1ki_1}\beta^*_{2ki_2}\beta^*_{3ki_3}\Big)^2.$$

Following the same strategy above, we obtain

$$\begin{aligned}
\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2 \le & 3(1+c)\Big(\sum_{k=1}^{K}\|\bar{\boldsymbol\beta}_{1k}-\bar{\boldsymbol\beta}_{1k}^*\|_2^2+\sum_{k=1}^{K}\|\bar{\boldsymbol\beta}_{2k}-\bar{\boldsymbol\beta}_{2k}^*\|_2^2\\
& +\sum_{k=1}^{K}\|\bar{\boldsymbol\beta}_{3k}-\bar{\boldsymbol\beta}_{3k}^*\|_2^2\Big)\Big(\sum_{k=1}^{K}(\sqrt[3]{\eta_k^*})^4\Big).
\end{aligned}$$

This ends the proof. ■

3.9.12 Proof of Lemma 13

First of all, let us state a lemma to illustrate the effect of weight ϕ.

Lemma 15 . Consider $\{y_i\}_{i=1}^n$ come from either non-symmetric tensor estimation model (3.5.1) or symmetric tensor estimation model (3.2.1). Suppose Conditions 3-5 hold. Then $\phi=\frac{1}{n}\sum_{i=1}^n y_i^2$ is upper and lower bounded by

$$(16-6\Gamma^3-9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2\le\frac{1}{n}\sum_{i=1}^{n}y_i^2\le(16+6\Gamma^3+9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2,$$

with probability at least $1-(K^2+K+3)/n$, where Γ is the incoherence parameter.

According to Lemma 15, $\frac{1}{n}\sum_{i=1}^n y_i^2$ approximates $(\sum_{k=1}^K \eta_k^*)^2$ up to some constants with high probability. Moreover, we know that from (3.9.5), $\max_k |\eta_k - \eta_k^*| \le \varepsilon_0$ for some small ε_0. Based on those two facts described above, we replace η_k by η_k^* and ϕ by $(\sum_{k=1}^K \eta_k^*)^2$ for the sake of completeness. Note that this change could only result in some constant scale changes for final results. Similar simplification was used in matrix recovery scenario (101). Therefore, we define the weighted estimator and weighted true parameter as $\bar{\boldsymbol{\beta}}_k = \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k$, $\bar{\boldsymbol{\beta}}_k^* = \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*$. Correspondingly, define the gradient function $\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)$ on F as

$$\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)_F = \frac{6\sqrt[3]{\eta_k^*}}{n}\sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'})^3 - y_i\Big)(\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2 \boldsymbol{x}_{i_F},$$

and its noiseless version as

$$\nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)_F = \frac{6\sqrt[3]{\eta_k^*}}{n}\sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'})^3 - \sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3\Big)(\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2 \boldsymbol{x}_{i_F}. \tag{3.9.30}$$

According to the definition of thresholding function (3.2.8), $\widetilde{\boldsymbol{\beta}}_k^+$ can be written as

$$\widetilde{\boldsymbol{\beta}}_k^+ = \boldsymbol{\beta}_k - \frac{\mu}{\phi}\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)_F + \frac{\mu}{\phi} h(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\gamma}_k,$$

where $\boldsymbol{\gamma}_k \in \mathbb{R}^p$ satisfies $\mathrm{supp}(\boldsymbol{\gamma}_k) \subset F$, $\|\boldsymbol{\gamma}_k\|_\infty \le 1$ and $h(\bar{\boldsymbol{\beta}}_k)$ is defined as

$$h(\bar{\boldsymbol{\beta}}_k) = \frac{\sqrt{4\log(np)}}{n}\sqrt{\sum_{i=1}^n \Big(\sum_{k=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^3 - y_i\Big)^2 \eta_k^{*\frac{2}{3}} (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2}. \tag{3.9.31}$$

Moreover, we denote $\boldsymbol{z}_k = \bar{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}}_k^*$. With a little abuse of notations, we also drop the subscript F in this section for notation simplicities.

We expand and decompose the sum of square error by three parts as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k^*}\widetilde{\boldsymbol{\beta}}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
&=\sum_{k=1}^{K}\left\|\boldsymbol{z}_k-\frac{\mu\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)+\frac{\mu\sqrt[3]{\eta_k^*}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\gamma}_k\right\|_2^2\\
&=\underbrace{\sum_{k=1}^{K}\left\|\boldsymbol{z}_k-\mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\right\|_2^2}_{\text{A: gradient update effect}}+\underbrace{\sum_{k=1}^{K}\left\|\frac{\mu\sqrt[3]{\eta_k^*}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\gamma_k\right\|_2^2}_{\text{B: threshoding effect}}\\
&+\underbrace{\sum_{k=1}^{K}\left\langle\boldsymbol{z}_k-\mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k),\frac{\mu\sqrt[3]{\eta_k}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\gamma_k\right\rangle}_{\text{C: cross term}}.
\end{aligned}\tag{3.9.32}$$

In the following proof, we will bound three parts sequentially.

Bounding gradient update effect In order to separate the optimization error and statistical error, we use the noiseless gradient $\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)$ as a bridge such that A can be decomposed as

$$\begin{aligned}
A=&\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2-2\mu\sum_{k=1}^{K}\left\langle\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k),\boldsymbol{z}_k\right\rangle+\mu^2\sum_{k=1}^{K}\left\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\right\|_2^2\\
\le&\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2-2\mu\underbrace{\sum_{k=1}^{K}\left\langle\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k),\boldsymbol{z}_k\right\rangle}_{A_1}+2\mu^2\underbrace{\sum_{k=1}^{K}\left\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)\right\|_2^2}_{A_2}\\
&+2\mu^2\underbrace{\sum_{k=1}^{K}\left\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\Big(\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)-\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big)\right\|_2^2}_{A_3}\\
&+2\mu\underbrace{\sum_{k=1}^{K}\left\langle\boldsymbol{z}_k,\frac{\sqrt[3]{\eta_k^*}}{\phi}\Big(\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)-\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big)\right\rangle}_{A_4},
\end{aligned}\tag{3.9.33}$$

where A_1 and A_2 quantify the optimization error, A_3 quantifies the statistical error, and A_4 is a cross term which can be negligible comparing with the rate of the statistical error. The lower bound for A_1 and upper bound for A_2 together coincide with the verification of regularity conditions in the matrix recovery case (92).

Step One: Lower bound for A_1. Plugging in $\phi = (\sum_{k=1}^K \eta_k^*)^2$, we have

$$K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}} \leq \frac{(\sqrt[3]{\eta_k^*})^2}{\phi} = \frac{(\sqrt[3]{\eta_k^*})^2}{(\sum_{k=1}^K \eta_k^*)^2} \leq K^{-2}R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}. \tag{3.9.34}$$

According to the definition of noiseless gradient $\nabla_k \widetilde{\mathcal{L}}(\boldsymbol{\beta}_k)$ and $\boldsymbol{z}_k$, A_1 can be expanded and decomposed sequentially by nine terms,

$$\begin{aligned}
A_1 \geq & K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}\Big[\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow \ A_{11} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow \ A_{12} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^3\Big) \Leftarrow \ A_{13} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow \ A_{14} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow \ A_{15} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow \ A_{16} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow \ A_{17} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow \ A_{18} \\
& + \frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^3\Big)\Big] \Leftarrow \ A_{19},
\end{aligned} \tag{3.9.35}$$

where A_{11} is the main term according to the order of $\bar{\boldsymbol{\beta}}_k^*$, while A_{12} to A_{19} are remainder terms. The proof of lower bound for A_{11} to A_{19} follows two steps:

1. Calculate and lower bound the expectation of each term through Lemma 3.12.1: high-order Gaussian moment;

2. Argue that the empirical version is concentrated around their expectation with high probability through Lemma 1: high-order concentration inequality.

Bounding A_{11}. Note that A_{11} involves the product of dependent Gaussian vectors. This brings difficulties on both the calculation of expectations and the use of concentration inequality. According to the high-order Gaussian moment results in Lemma 3.12.1, the expectation of A_{11} can be calculated explicitly as

$$\begin{aligned}\mathbb{E}(A_{11}) = 36\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k}^{*})^2(\boldsymbol{z}_{k'}^{\top}\boldsymbol{z}_k) \Leftarrow\ I_1 \\ + 72\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k}^{*})(\boldsymbol{z}_{k'}^{\top}\bar{\boldsymbol{\beta}}_{k}^{*})(\boldsymbol{z}_{k}^{\top}\bar{\boldsymbol{\beta}}_{k'}^{*}) \Leftarrow\ I_2 \\ + 108\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k}^{*})(\boldsymbol{z}_{k'}^{\top}\bar{\boldsymbol{\beta}}_{k'}^{*})(\boldsymbol{z}_{k}^{\top}\bar{\boldsymbol{\beta}}_{k}^{*}) \Leftarrow\ I_3 \\ + 54\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k'}^{*})(\bar{\boldsymbol{\beta}}_{k}^{*\top}\bar{\boldsymbol{\beta}}_{k}^{*})(\boldsymbol{z}_{k'}^{\top}\boldsymbol{z}_k) \Leftarrow\ I_4.\end{aligned} \tag{3.9.36}$$

Note that I_1 to I_4 involve the summation of K^2 term. To use incoherence Condition 3, we isolate K terms with $k = k'$. Then, I_1 to I_4 could be lower bounded as

$$\begin{aligned} I_1 &\geq 36\eta_{\min}^{*4/3}\Big[\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2 - \Gamma^2\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)^2\Big] \\ I_2 &\geq 72\eta_{\min}^{*4/3}\Big[\sum_{k=1}^{K}(\boldsymbol{z}_k^{\top}\bar{\boldsymbol{\beta}}_{k}^{*})^2 - \Gamma\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)^2\Big] \\ I_3 &\geq 108\eta_{\min}^{*4/3}\Big[\sum_{k=1}^{K}(\boldsymbol{z}_k^{\top}\bar{\boldsymbol{\beta}}_{k}^{*})^2 - \Gamma\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)^2\Big] \\ I_4 &\geq 54\eta_{\min}^{*4/3}\Big\|\sum_{k=1}^{K}\boldsymbol{z}_k\Big\|_2^2 \geq 0, \end{aligned}$$

where Γ is the incoherence parameter. Putting the above four bounds together, they jointly provide

$$\mathbb{E}(A_{11}) \geq 36\eta_{\min}^{*4/3}\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2 - \Big(36\eta_{\min}^{*4/3}\Gamma^2 + 180\eta_{\min}^{*4/3}\Gamma\Big)\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)^2. \tag{3.9.37}$$

On the other hand, repeatedly using Lemma 1, we obtain that with probability at least $1-1/n$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\Big((\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2-\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big)\Big|$$
$$\leq C\frac{(\log n)^3}{\sqrt{n}}(\sqrt[3]{\eta_{\max}^*})^4\|\boldsymbol{z}_{k'}\|_2\|\boldsymbol{z}_k\|_2.$$

Taking the summation over $k,k'\in[K]$, it could further imply that for some absolute constant C,

$$\Big|A_{11}-\mathbb{E}(A_{11})\Big|\leq 18C\frac{(\log n)^3}{\sqrt{n}}(\sqrt[3]{\eta_{\max}^*})^4\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)^2, \tag{3.9.38}$$

with probability at least $1-K^2/n$. Combining (3.9.37) and (3.9.38), we obtain with probability at least $1-K^2/n$,

$$\begin{aligned}&K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}A_{11}\\ \geq&\Big[36K^{-2}R^{-\frac{8}{3}}-K^{-\frac{3}{2}}\Big(216R^{-\frac{8}{3}}\Gamma+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big]\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2,\end{aligned} \tag{3.9.39}$$

where $R=\eta_{\max}^*/\eta_{\min}^*$. Here, we use the fact $\Gamma\leq 1$ and $(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2)^2\leq K(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2)$.

Bounding A_{12} to A_{19}: For remainder terms, we follow the same proof strategy. According to Lemma 3.12.1, the expectation of A_{12} can be calculated as

$$\begin{aligned}\mathbb{E}(A_{12}) \;=\; &36\sum_{k=1}^{K}\sum_{k'=1}^{K}(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'}^{*\top})^2(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*)\Leftarrow\; I_1\\ &+72\sum_{k=1}^{K}\sum_{k'=1}^{K}(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'}^*)(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\boldsymbol{z}_k)\Leftarrow\; I_2\\ &+108\sum_{k=1}^{K}\sum_{k'=1}^{K}(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'})(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_k^*)\Leftarrow\; I_3\\ &+54\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_k^\top\boldsymbol{z}_k)\Leftarrow\; I_4.\end{aligned}$$

Let us analyze I_1 first. Under (3.9.5), $\|\boldsymbol{z}_k\|_2 \le \varepsilon_0 \sqrt[3]{\eta_k^*}$, it suffices to show that

$$\begin{aligned}\sum_{k=1}^K\sum_{k'=1}^K(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'})^2(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*) &\ge -\sum_{k=1}^K\sum_{k'=1}^K\|\boldsymbol{z}_k\|_2^2\|\bar{\boldsymbol{\beta}}_{k'}^*\|_2^2\|\boldsymbol{z}_k'\|_2\|\bar{\boldsymbol{\beta}}_k^*\|_2\\ &\ge -\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2.\end{aligned}$$

This immediately implies a lower bound for $\mathbb{E}(A_{12})$ after we bound similarly for I_2, I_3 and I_4,

$$\mathbb{E}(A_{12}) \ge -270\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2. \tag{3.9.40}$$

By Lemma 1, we obtain for some absolute constant C,

$$\begin{aligned}&K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}A_{12}\\ \ge&K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}\Big[\mathbb{E}(A_{12}) - 18C\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2\frac{(\log n)^3}{\sqrt{n}}\Big]\\ \ge& -K^{-1}R^{-\frac{2}{3}}\varepsilon_0\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big),\end{aligned} \tag{3.9.41}$$

with probability at least $1-K^2/n$. The detail derivation is the same as in (3.9.39), so we omit here.

Similarly, the lower bounds of A_{13} to A_{19} can be derived as follows

$$\begin{aligned}&K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{14} \ge -K^{\frac{1}{2}}\varepsilon_0\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{13}, A_{15}, A_{17} \ge -K^{\frac{1}{2}}\varepsilon_0^2\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{16}, A_{18} \ge -K^{\frac{1}{2}}\varepsilon_0^3\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{19} \ge -K^{\frac{1}{2}}\varepsilon_0^4\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big).\end{aligned} \tag{3.9.42}$$

Putting (3.9.39), (3.9.41) and (3.9.42) together, we have with probability at least $1 - 9K^2/n$,

$$\begin{aligned} A_1 \quad \geq \quad & \Big[36K^{-2}R^{-\frac{8}{3}} - K^{-\frac{3}{2}}\Big(2160R^{-\frac{3}{3}}\Gamma + 18C\frac{(\log n)^3}{\sqrt{n}}\Big) \\ & -8\varepsilon_0 K^{-1}R^{-\frac{2}{3}}\Big(270 + 18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big). \end{aligned}$$

For the above bound,

- When the sample size satisfies $n \geq (18CK^{1/2}R^{8/3}(\log n)^3)^2$, we have

$$\max\Big\{18K^{-\frac{3}{2}}C\frac{(\log n)^3}{\sqrt{n}}, 8\varepsilon_0 K^{-1}R^{-\frac{2}{3}}18C\frac{(\log n)^3}{\sqrt{n}}\Big\} \leq K^{-2}R^{-\frac{8}{3}}.$$

- When $\varepsilon_0 \leq K^{-1}R^{-2}/2160$, we have

$$8\varepsilon_0 K^{-1}R^{-\frac{2}{3}}270 \leq K^{-2}R^{-\frac{8}{3}}.$$

- When the incoherence parameter satisfies $\Gamma \leq K^{-1/2}/216$, we have

$$K^{-\frac{3}{2}}2160R^{-\frac{8}{3}}\Gamma \leq K^{-2}R^{-\frac{8}{3}}.$$

Note that those above conditions can be fulfilled by Conditions 3, 5 and (3.9.5). Thus, we are able to simplify A_1 by

$$A_1 \geq 32K^{-2}R^{-\frac{8}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big), \tag{3.9.43}$$

with probability at least $1 - 9K^2/n$.

Step Two: Upper bound for A_2. We observe the fact that

$$\begin{aligned} A_2 &= \sum_{k=1}^{K}\Big\|\frac{1}{\phi}\sqrt[3]{\eta_k^*}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2 \\ &= \sup_{\boldsymbol{w}\in\mathbb{S}^{Ks-1}}\Big|\Big\langle\sum_{k=1}^{K}\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{w}\Big\rangle\Big|^2, \end{aligned} \tag{3.9.44}$$

where $\mathbb{S}$ is a unit sphere. It is equivalent to show for any $\boldsymbol{w} \in \mathbb{S}^{Ks-1}$, $A_2' = |\langle \sum_{k=1}^K \frac{\sqrt[3]{\eta_k^*}}{\phi} \nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{w} \rangle|$ is upper bounded. According to the definition of noiseless gradient (3.9.30), A_2' is explicitly written as

$$A_2' = \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 - \sum_{k'=1}^K (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3 \Big) \Big(\sum_{k=1}^K \frac{(\sqrt[3]{\eta_k^*})^2}{\phi} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2 (\boldsymbol{x}_i^\top \boldsymbol{w}) \Big).$$

Following by (3.9.34) and (3.9.35), similar decomposition can be made for A_2' as follows, where the only difference is that we replace one $\boldsymbol{x}_i^\top \boldsymbol{z}_k$ by $\boldsymbol{x}_i^\top \boldsymbol{w}$.

$$\begin{aligned}
A_2' \;\leq\; & K^{-2} R^{\frac{2}{3}} \eta_{\min}^{*-\frac{4}{3}} \Big[\frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2 \Big) \Leftarrow A_{21}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*) \Big) \Leftarrow A_{22}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^2 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2 (\boldsymbol{x}_i^\top \boldsymbol{w}) \Big) \Leftarrow A_{23}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2 \Big) \Leftarrow A_{24}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*) \Big) \Leftarrow A_{25}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}) \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*) \Big) \Leftarrow A_{26}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2 \Big) \Leftarrow A_{27}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K 2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)(\boldsymbol{x}_i^\top \boldsymbol{w})(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*) \Big) \Leftarrow A_{28}' \\
& + \frac{6}{n} \sum_{i=1}^n \Big(\sum_{k'=1}^K 3(\boldsymbol{x}_i^\top \boldsymbol{z}_{k'})^3 \sum_{k=1}^K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2 (\boldsymbol{x}_i^\top \boldsymbol{w}) \Big) \Big]. \Leftarrow A_{29}'
\end{aligned}$$

Let's bound A'_{21} first. By using the same technique when calculating $\mathbb{E}(A_{11})$ in (3.9.36), we derive an upper bound for $\mathbb{E}(A'_{21})$,

$$\begin{aligned}\mathbb{E}(A'_{21}) &\leq 36\eta_{\max}^{*\frac{4}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2+(K-1)\sum_{k=1}^{K}\Gamma\|\boldsymbol{z}_k\|_2\Big)\\ &+ 180\eta_{\max}^{*\frac{4}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2+(K-1)\sum_{k=1}^{K}\Gamma\|\boldsymbol{z}_k\|_2\Big)+54\eta_{\max}^{*\frac{4}{3}}\Big(K\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big).\end{aligned}$$

Equipped with Lemma 2 and the definition of tensor spectral norm (3.1.3), it suffices to bound A'_{21} by

$$R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}K^{-\frac{1}{2}}A'_{21}\leq K^{-2}R^2\Big[216+54K+216K\Gamma+18CK\delta_{n,p,s}\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)$$

with probability at least $1-10K^2/n^3$, where $\delta_{n,p,s}$ is defined in (3.3.7).

The upper bounds for A'_{22} to A'_{29} follow similar forms. Combining them together, we can derive an upper bound for A'_2 as follows

$$\begin{aligned}A'_2 &\leq K^{-2}R^2\Big[216+270K+18CK\delta_{n,p,s}\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)\\ &\leq K^{-2}R^2\Big[220+270K\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big),\end{aligned}$$

with probability at least $1-90K^2/n^3$, where the second inequality utilizes Condition 5. Therefore, the upper bound of A_2 is given as follows

$$A_2\leq K^{-1}R^4[220+270K]^2\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big), \tag{3.9.45}$$

with probability at least $1-90K^2/n^3$.

Step Three: Upper bound for A_3. By the definition of noisy gradient and noiseless gradient, A_3 is explicitly written as

$$\begin{aligned}A_3 &= \sum_{k=1}^{K}\Big\|\frac{(\sqrt[3]{\eta_k^*})^2}{\phi}\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2\boldsymbol{x}_i\Big\|_2^2\\ &\leq K^{-4}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\sum_{k=1}^{K}\Big(\sqrt{Ks}\max_j\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}\Big)^2,\end{aligned}$$

where the second inequality comes from (3.9.34). For fixed $\{\epsilon_i\}_{i=1}^n$, applying Lemma 1, we have

$$\Big|\sum_{i=1}^n \epsilon_i(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2 x_{ij} - \mathbb{E}\Big(\sum_{i=1}^n \epsilon_i(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2 x_{ij}\Big)\Big| \le C(\log n)^{\frac{3}{2}}\|\boldsymbol{\epsilon}\|_2\|\bar{\boldsymbol{\beta}}_k\|_2^2,$$

with probability at least $1-1/n$. Together with Lemma 23, we obtain for any $j \in [Ks]$,

$$\Big|\frac{6}{n}\sum_{i=1}^n \epsilon_i(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2 x_{ij}\Big| \le 6CC_0\sigma\|\bar{\boldsymbol{\beta}}_k\|_2^2\frac{(\log n)^{3/2}}{\sqrt{n}},$$

with probability at least $1-4/n$, where σ is the noise level. According to (3.9.5),

$$\Big\|\bar{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}}_k^*\Big\|_2^2 \le \sum_{k=1}^K \Big\|\bar{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}}_k^*\Big\|_2^2 \le K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2,$$

which further implies $\|\bar{\boldsymbol{\beta}}_k\|_2^2 \le (1+K^{\frac{1}{2}}\varepsilon_0)^2\eta_{\max}^{*\frac{2}{3}}$. Equipped with union bound over $j \in [Ks]$,

$$\max_{j\in[Ks]}\Big|\frac{6}{n}\sum_{i=1}^n \epsilon_i(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2 x_{ij}\Big| \le 6CC_0\sigma(1+K^{\frac{1}{2}}\varepsilon_0)^2(\sqrt[3]{\eta_{\max}^*})^2\frac{(\log n)^{3/2}}{\sqrt{n}},$$

with probability at least $1-4Ks/n$. Letting $C = 6C_0(Ce)^{-2/3}(1+K^{\frac{1}{2}}\varepsilon_0)^2$,

$$A_3 \le C\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}\sigma^2K^{-2}\frac{s(\log n)^3}{n}, \tag{3.9.46}$$

with probability at least $1-4Ks/n$.

Step Four: Upper bound for A_4. This cross term can be written as

$$A_4 = 2\sum_{k=1}^K \frac{\mu}{\phi}(\sqrt[3]{\eta_k^*})^2\Big(\frac{1}{n}\sum_{i=1}^n \epsilon_i(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^2(\boldsymbol{x}_i^\top \boldsymbol{z}_k)\Big).$$

To bound this term, we take the same step in Step Three which fixes the noise term $\{\epsilon_i\}_{i=1}^n$ first. Similarly, we obtain with probability at least $1-4K/n$,

$$A_4 \le 2C\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}K^{-1}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{2}{3}}. \tag{3.9.47}$$

This term is negligible in terms of the order when comparing with (3.9.46).

Summary. Putting the bounds (3.9.43), (3.9.45), (3.9.46) and (3.9.47) together, we achieve an upper bound for gradient update effect as follows,

$$\begin{aligned} A \leq & \Big(1 - 64\mu K^{-2} R^{-\frac{8}{3}} + 2\mu^2 K^{-1} R^4 [220 + 270K]^2\Big) \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2 \\ & + 4\mu C K^{-2} \eta_{\min}^{*-\frac{4}{3}} R^{\frac{8}{3}} \frac{\sigma^2 s (\log n)^3}{n}, \end{aligned} \tag{3.9.48}$$

with probability at least $1 - (18K^2 + 4K + 4Ks)/n$. ■

Bounding thresholding effect The thresholding effect term in (3.9.32) can also be decomposed into optimization error and statistical error. Recall that B can be explicitly written as

$$B = \sum_{k=1}^{K} \Big\| \mu \frac{\eta_k^{*\frac{2}{3}}}{\phi} \frac{4\sqrt{\log(np)}}{n} \sqrt{\sum_{i=1}^{n} \Big(\sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 - y_i \Big)^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4} \gamma_k \Big\|_2^2,$$

where $\mathrm{supp}(\gamma_k) \subset F_k$ and $\|\gamma_k\|_\infty \leq 1$. By using $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\begin{aligned} B \leq & \mu^2 \frac{64Ks \log p}{n} \Big[\underbrace{\frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 - \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3 \Big) \Big(\sum_{k=1}^{K} \frac{\eta_k^{*\frac{4}{3}}}{\phi^2} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \Big)}_{B_1\text{:optimization error}} \\ & + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \sum_{k=1}^{K} \frac{\eta_k^{*\frac{4}{3}}}{\phi^2} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4}_{B_2\text{:statistical error}} \Big]. \end{aligned}$$

Bounding B_1. This optimization error term shares similar structure with (3.9.44)

but with higher order. Therefore, we follow the same idea as we did in bounding (3.9.44). Following by (3.9.34) and some basic expansions and inequalities,

$$\begin{aligned}
B_1 \leq & K^{-2}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\frac{1}{n}\Big(\sum_{k'=1}^{K}(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^3-\sum_{k'=1}^{K}(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^3\Big)\Big(\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4\Big)\\
\leq & K^{-2}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{K}3K(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^6+9K(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^4(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\\
& +9K(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^4\Big)\sum_{k'=1}^{K}(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^4\Big].
\end{aligned}$$

The main term is $(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^4$ according to the order of $\bar{\boldsymbol{\beta}}_k^*$. We bound the main term first. Note that there exists some positive large constant C such that

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^4(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^4\Big)\leq C\|\boldsymbol{z}_k\|_2^2\|\bar{\boldsymbol{\beta}}_k^*\|_2^4\|\bar{\boldsymbol{\beta}}_{k'}\|_2^4.$$

Together with Lemma 1 and (3.9.5), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{k'=1}^{K}\Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^4(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^4\Big)\\
\leq & C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)K^2\eta_{\max}^{*\frac{8}{3}}(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2.
\end{aligned}$$

with probability at least $1-3K^2/n$. Overall, the upper bound of B_1 takes the form

$$\begin{aligned}
B_1 \leq & K^{-2}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\Big[18C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)K^2\eta_{\max}^{*\frac{8}{3}}(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big]\\
\leq & R^4 18C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2,
\end{aligned}\tag{3.9.49}$$

with probability at least $1-3K^2/n$.

Bounding B_2. We rewrite B_2 by

$$B_2=\sum_{k=1}^{K}\frac{\eta_k^{*\frac{4}{3}}}{\phi^2}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4\Big).$$

For fixed $\{\epsilon_i\}_{i=1}^n$, accordingly to Lemma 1, we have

$$\Big|\sum_{i=1}^n \epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4 - \mathbb{E}\Big(\sum_{i=1}^n \epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4\Big)\Big| \le C(\log n)^2\|\boldsymbol{\epsilon}^2\|_2\|\bar{\boldsymbol{\beta}}_k\|_2^4.$$

Note that $\mathbb{E}((\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4) = 3\|\bar{\boldsymbol{\beta}}_k\|_2^4$. It will reduce to

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4 \le \Big(\frac{3}{n}\sum_{i=1}^n \epsilon_i^2 + C\frac{(\log n)^2}{n}\|\boldsymbol{\epsilon}^2\|_2\Big)\|\bar{\boldsymbol{\beta}}_k\|_2^4.$$

From Lemma 23, with probability at least $1-3/n$,

$$|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2| \le C_0\sigma^2,\ \ \frac{1}{n}\|\boldsymbol{\epsilon}^2\|_2 \le C_0\frac{\sigma^2}{\sqrt{n}}.$$

Combining the above two inequalities, we obtain

$$\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4\Big| \le 6C_0\sigma^2\|\bar{\boldsymbol{\beta}}_k\|_2^4, \tag{3.9.50}$$

with probability at least $1-7/n$. Plugging in the definition of ϕ and (3.9.5), B_2 is upper bounded by

$$B_2 \le 6C_0\sigma^2(1+K^{\frac{1}{2}}\varepsilon_0)^4\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}K^{-3}, \tag{3.9.51}$$

with probability at least $1-7K/n$.

Summary. Putting the bounds (3.9.49) and (3.9.51) together, we have similar upper bound for thresholded effect,

$$B \le C_2\mu^2R^4\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2 + C_3\mu^2\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}K^{-2}\frac{\sigma^2 s\log p}{n}, \tag{3.9.52}$$

with probability at least $1-(3K^2+7K)/n$. ■

Ensemble From the definition of γ_k, it's not hard to see actually the cross term C is equal to zero. Combining the upper bound of gradient update effect (3.9.48) and thresholding effect (3.9.52) together, we obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
\leq&\Big(1-64\mu K^{-2}R^{-\frac{8}{3}}+3\mu^2K^{-1}R^4[220+270K]^2\Big)\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big)\\
&+2C_3\mu^2R^{\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n}.
\end{aligned}$$

As long as the step size μ satisfies

$$0<\mu\leq\frac{32R^{-20/3}}{3K[220+270K]^2},$$

we reach the conclusion

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
\leq&\Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
&+2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n},
\end{aligned}\tag{3.9.53}$$

with probability at least $1-4Ks/n$. ■

3.9.13 Proof of Lemma 14

Let us consider k-th component first. Without loss of generality, suppose $F\subset\{1,2,\ldots,Ks\}$. For $j=Ks+1,\ldots,p$,

$$\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)=\frac{2}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i\Big)\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2x_{ij},\tag{3.9.54}$$

and it's not hard to see the independence between $\{\boldsymbol{x}_i^\top\boldsymbol{\beta}_k,y_i\}$ and x_{ij}. Applying standard Hoeffding's inequality, we have with probability at least $1-\frac{1}{n^2p^2}$,

$$\Big|\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)\Big|\leq\frac{\sqrt{4\log(np)}}{n}\sqrt{\sum_{i=1}^{n}(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i)^2(\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k))^2}=h(\boldsymbol{\beta}_k).$$

Equipped with union bound, with probability at least $1-\frac{1}{n^2p}$,

$$\max_{Ks+1\le j\le p}\Big|\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)\Big|\le h(\boldsymbol{\beta}_k).$$

Therefore, according to the definition of thresholding function $\varphi(\boldsymbol{x})$, we obtain the following equivalence,

$$\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)\Big)=\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)_F\Big),\tag{3.9.55}$$

holds for $k\in[K]$, with probability at least $1-\frac{1}{n^2p}$. (3.9.55) also provides that $\mathrm{supp}(\boldsymbol{\beta}_k^+)\subset F$ for every $k\in[K]$, which further implies $F^+\subset F$. Now we end the proof. ■

3.9.14 Proof of Lemma 15

First, we consider symmetric case. According to the definition of $\{y_i\}_{i=1}^n$ from symmetric tensor estimation model (3.2.1), we separate the random noise ϵ_i by the following expansion,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n y_i^2 &= \frac{1}{n}\sum_{i=1}^n\Big[\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3+\epsilon_i\Big]^2\\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n(\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2}_{I_1}+\underbrace{\frac{2}{n}\sum_{i=1}^n\epsilon_i\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3}_{I_2}+\underbrace{\frac{1}{n}\sum_{i=1}^n\epsilon_i^2}_{I_3}\end{aligned}\tag{3.9.56}$$

Bounding I_1. We expand i-th component of I_1 as follows

$$\begin{aligned}&(\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2\\ =&\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6+2\sum_{k_i<k_j}\eta_{k_i}^*\eta_{k_j}^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3.\end{aligned}\tag{3.9.57}$$

As shown in Corollary 3.12.1, the expectations of above two parts takes forms of

$$\begin{aligned}&\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3=6(\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*)^3+9(\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*)\|\boldsymbol{\beta}_{k_i}^*\|_2^2\|\boldsymbol{\beta}_{k_j}^*\|_2^2\\ &\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6=15\|\boldsymbol{\beta}_k^*\|_2^2.\end{aligned}$$

Recall that $\|\boldsymbol{\beta}_k^*\|_2 = 1$ for any $k \in [K]$ and Condition 3 implies for any $k_i \neq k_j$, $|\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*| \leq \Gamma$, where Γ is the incoherence parameter. Thus, $\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3$ is upper bounded by

$$\left|\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3\right| \leq 6\Gamma^3 + 9\Gamma, \text{ for any } k_i \neq k_j. \tag{3.9.58}$$

By using the concentration result in Lemma 1, we have with probability at least $1 - 1/n$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6 - \mathbb{E}(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6)\right| \leq C_1\frac{(\log n)^3}{\sqrt{n}},\\
&\left|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3 - \mathbb{E}(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3)\right| \leq C_1\frac{(\log n)^3}{\sqrt{n}}.
\end{aligned} \tag{3.9.59}$$

Putting (3.9.57),(3.9.58) and (3.9.59) together, this essentially provides an upper bound for I_1, namely

$$\frac{1}{n}\sum_{i=1}^n(\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2 \leq \left(15 + 6\Gamma^3 + 9\Gamma + 2C_1\frac{(\log n)^3}{\sqrt{n}}\right)(\sum_{k=1}^K\eta_k^*)^2, \tag{3.9.60}$$

with probability at least $1 - K^2/n$.

Bounding I_2. Since the random noise $\{\epsilon_i\}_{i=1}^n$ is of mean zero and independent of $\{\boldsymbol{x}_i\}$, we have

$$\mathbb{E}(\epsilon_i\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3) = 0.$$

By using the independence and Corollary 1, we have

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3 \geq C_2\frac{(\log n)^{\frac{3}{2}}}{n}\sqrt{n}\sigma\Big)\\
&\leq \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3 \geq C_2\frac{(\log n)^{\frac{3}{2}}}{n}\sqrt{n}\sigma\Big|\|\boldsymbol{\epsilon}\|_2 \leq C_0\sigma\sqrt{n}\Big) + \mathbb{P}\Big(\|\boldsymbol{\epsilon}\|_2 \geq C_0\sqrt{n}\sigma\Big)\\
&\leq \frac{1}{n} + \frac{3}{n} = \frac{4}{n}.
\end{aligned}$$

This further implies that

$$\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3\epsilon_i \leq (\sum_{k=1}^K\eta_k^*)C_2\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}\sigma, \tag{3.9.61}$$

with probability at least $1 - 4K/n$.

Bounding I_3. As shown in Lemma 23, the random noise ϵ_i with sub-exponential tail satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \leq C_3\sigma^2. \tag{3.9.62}$$

with probability at least $1 - 3/n$.

Overall, putting (3.9.60), (3.9.61) and (3.9.62) together, we have with probability at least $1 - (K^2 + 4K + 3)/n$,

$$\frac{\frac{1}{n}\sum_{i=1}^{n} y_i^2}{(\sum_{k=1}^{K}\eta_k^*)^2} \leq 15 + 6\Gamma^3 + 9\Gamma + 2C_1\frac{(\log n)^3}{\sqrt{n}} + \frac{2C_2\sigma}{(\sum_{k=1}^{K}\eta_k^*)}\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}} + \frac{C_3\sigma^2}{(\sum_{k=1}^{K}\eta_k^*)^2}.$$

Under Conditions 4 & 5, the above bound reduces to

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 \leq (16 + 6\Gamma^3 + 9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2,$$

with probability at least $1 - (K^2 + 4K + 3)/n$. The proof of lower bound is similar, and hence is omitted here.

Similar results will also hold for non-symmetric tensor estimation model. Throughout the proof, the only difference is that

$$\mathbb{E}(\boldsymbol{u}_i^\top\boldsymbol{\beta}_{1k}^*)^2(\boldsymbol{v}_i^\top\boldsymbol{\beta}_{2k}^*)^2(\boldsymbol{w}_i^\top\boldsymbol{\beta}_{3k}^*)^2 = 1.$$

■

3.10 Non-symmetric Tensor Estimation

3.10.1 Conditions and Algorithm

In this subsection, we provide several essential conditions for Theorem 7 and the detail algorithm for non-symmetric tensor estimation.

Condition 6 (Uniqueness of CP-decomposition). The CP-decomposition form (3.5.2) is unique in the sense that if there exists another CP-decomposition $\mathscr{T}^* = \sum_{k=1}^{K'} \eta_k^{*'} \boldsymbol{\beta}_{1k}^{*'} \circ \boldsymbol{\beta}_{2k}^{*'} \circ \boldsymbol{\beta}_{3k}^{*'}$, it must have $K = K'$ and be invariant up to a permutation of $\{1, \ldots, K\}$.

Condition 7 (Parameter space). The CP-decomposition of $\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*$ satisfies

$$\|\mathscr{T}^*\|_{op} \leq C_1 \eta_{\max}^*, \quad K = \mathcal{O}(s), \quad \text{and} \quad R = \eta_{\max}^*/\eta_{\min}^* \leq C_2$$

for some absolute constants C_1, C_2.

Condition 8 (Parameter incoherence). The true tensor components are incoherent such that

$$\Gamma := \max_{k_i \neq k_j} \Big\{ |\langle \boldsymbol{\beta}_{1k_i}^*, \boldsymbol{\beta}_{1k_j}^* \rangle|, |\langle \boldsymbol{\beta}_{2k_i}^*, \boldsymbol{\beta}_{2k_j}^* \rangle|, |\langle \boldsymbol{\beta}_{3k_i}^*, \boldsymbol{\beta}_{3k_j}^* \rangle| \Big\} \leq C \min\{K^{-\frac{3}{4}} R^{-1}, s^{-\frac{1}{2}}\}.$$

Condition 9 (Random noise). We assume the random noise $\{\epsilon_i\}_{i=1}^n$ follows a sub-exponential tail with parameter σ satisfying $0 < \sigma < C \sum_{k=1}^{K} \eta_k^*$.

3.10.2 Proof of Theorem 7

The main distinguished part of the proof for non-symmetric update is Lemma 16: one-step oracle estimator, which is parallel to Lemma 11. For the sake of completeness, we limit our attention to rank-one case and only provide the theoretical development for one-step oracle estimator in this subsection. The generalization to general rank case follows the exact same idea in the proof of symmetric update by incorporating the incoherence condition (8).

For rank-one non-symmetric tensor estimation, the model (3.5.1) reduces to

$$y_i = \langle \eta^* \boldsymbol{\beta}_1^* \circ \boldsymbol{\beta}_2^* \circ \boldsymbol{\beta}_3^*, \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \rangle + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

Suppose $|\text{supp}(\boldsymbol{\beta}_1^*)| = s_1$, $|\text{supp}(\boldsymbol{\beta}_2^*)| = s_2$, $|\text{supp}(\boldsymbol{\beta}_3^*)| = s_3$ and denote $s = \max\{s_1, s_2, s_3\}$. Define $F_j^{(t)} = \text{supp}(\boldsymbol{\beta}_j^*) \cup \text{supp}(\boldsymbol{\beta}_j^{(t)})$, $F^{(t)} = \cup_{j=1}^3 F_j^{(t)}$ and the oracle estimator as

$$\widetilde{\boldsymbol{\beta}}_1^{(t+1)} = \varphi_{\frac{\mu}{\phi} h(\boldsymbol{\beta}_1^{(t)})} \Big(\boldsymbol{\beta}_j^{(t)} - \frac{\mu}{\phi} \nabla_1 \mathcal{L}(\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}, \boldsymbol{\beta}_3^{(t)})_{F^{(t)}} \Big),$$

Algorithm 4 Non-symmetric tensor estimation via cubic sketchings

Require: response $\{y_i\}_{i=1}^n$, sketching vector $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}_{i=1}^n$, truncation level d, step size μ, rank K, stopping error $\epsilon = 10^{-4}$.

1: **Step 1:** Calculate the moment-based tensor $\mathcal{T}$ as (3.5.4) and do sparse tensor decomposition on $\mathcal{T}$ to get a warm-start $\{\boldsymbol{\eta}^{(0)}, \boldsymbol{B}_1^{(0)}, \boldsymbol{B}_2^{(0)}, \boldsymbol{B}_3^{(0)}\}$.

2: **Step 2:** Let $t = 0$.

3: **Repeat** block-wise thresholded gradient update

4: • Compute threshold level $\boldsymbol{h}(\boldsymbol{B}_k)$ as defined in Step Two.

• Calculated block-wise thresholded gradient descent update

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{B}_1^{(t+1)}) &= \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B}_1)}{\phi}}\Big(\mathrm{vec}(\boldsymbol{B}_1^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1^{(t)}, \boldsymbol{B}_2^{(t)}, \boldsymbol{B}_3^{(t)})\Big) \\
\mathrm{vec}(\boldsymbol{B}_2^{(t+1)}) &= \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B}_2)}{\phi}}\Big(\mathrm{vec}(\boldsymbol{B}_2^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}_2}\mathcal{L}(\boldsymbol{B}_1^{(t)}, \boldsymbol{B}_2^{(t)}, \boldsymbol{B}_3^{(t)})\Big) \\
\mathrm{vec}(\boldsymbol{B}_3^{(t+1)}) &= \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B}_3)}{\phi}}\Big(\mathrm{vec}(\boldsymbol{B}_3^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}_3}\mathcal{L}(\boldsymbol{B}_1^{(t)}, \boldsymbol{B}_2^{(t)}, \boldsymbol{B}_3^{(t)})\Big),
\end{aligned}$$

where $\phi = \frac{1}{n}\sum_{i=1}^n y_i^2$. The detail form of $\nabla_{\boldsymbol{B}_1}\mathcal{L}, \nabla_{\boldsymbol{B}_2}\mathcal{L}, \nabla_{\boldsymbol{B}_3}\mathcal{L}$ can refer (3.5.5)

5: **Until** $\max\{\|\boldsymbol{B}_j^{(T+1)} - \boldsymbol{B}_j^{(T)}\|_F\} \le \epsilon$.

6: **Step 3:** Do column-wise normalization as

$$\begin{aligned}
\widehat{\boldsymbol{B}}_1 &= \Big(\frac{\boldsymbol{\beta}_{11}^{(T)}}{\|\boldsymbol{\beta}_{11}^{(T)}\|_2}, \ldots, \frac{\boldsymbol{\beta}_{1K}^{(T)}}{\|\boldsymbol{\beta}_{1K}^{(T)}\|_2}\Big), \\
\widehat{\boldsymbol{B}}_2 &= \Big(\frac{\boldsymbol{\beta}_{21}^{(T)}}{\|\boldsymbol{\beta}_{21}^{(T)}\|_2}, \ldots, \frac{\boldsymbol{\beta}_{2K}^{(T)}}{\|\boldsymbol{\beta}_{2K}^{(T)}\|_2}\Big), \\
\widehat{\boldsymbol{B}}_3 &= \Big(\frac{\boldsymbol{\beta}_{31}^{(T)}}{\|\boldsymbol{\beta}_{31}^{(T)}\|_2}, \ldots, \frac{\boldsymbol{\beta}_{3K}^{(T)}}{\|\boldsymbol{\beta}_{3K}^{(T)}\|_2}\Big).
\end{aligned}$$

And update the weight by

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}^{(0)} * (\|\boldsymbol{\beta}_{11}^{(T)}\|_2\|\boldsymbol{\beta}_{21}^{(T)}\|_2\|\boldsymbol{\beta}_{31}^{(T)}\|_2, \ldots, \|\boldsymbol{\beta}_{1K}^{(T)}\|_2\|\boldsymbol{\beta}_{3K}^{(T)}\|_2\|\boldsymbol{\beta}_{3K}^{(T)}\|_2)^\top.$$

The final estimator is $\widehat{\mathscr{T}} = \sum_{k=1}^K \widehat{\eta}_k \widehat{\boldsymbol{\beta}}_{1k} \circ \widehat{\boldsymbol{\beta}}_{2k} \circ \widehat{\boldsymbol{\beta}}_{3k}$.

7: **return** non-symmetric tensor estimator $\widehat{\mathscr{T}}$.

where $h(\boldsymbol{\beta}_1^{(t)})$ has the form of

$$\frac{\sqrt{4\log np}}{n}\sqrt{\sum_{i=1}^{n}\Big(\eta(\boldsymbol{u}_i^\top\boldsymbol{\beta}_1^{(t)})(\boldsymbol{v}_i^\top\boldsymbol{\beta}_2^{(t)})(\boldsymbol{w}_i^\top\boldsymbol{\beta}_3^{(t)})-y_i\Big)^2\eta^{\frac{2}{3}}(\boldsymbol{v}_i^\top\boldsymbol{\beta}_2^{(t)})^2(\boldsymbol{w}_i^\top\boldsymbol{\beta}_3^{(t)})^2}. \quad (3.10.1)$$

The definitions of $\widetilde{\boldsymbol{\beta}}_2^{(t+1)}$ and $\widetilde{\boldsymbol{\beta}}_3^{(t+1)}$ are similar.

Lemma 16 . Let $t \geq 0$ be an integer. Suppose Conditions 6-9 hold and $\{\boldsymbol{\beta}_j^{(t)}, \eta\}$ satisfies the following upper bound

$$\max_{j=1,2,3}\|\sqrt[3]{\eta}\boldsymbol{\beta}_j^{(t)}-\sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\|_2 \leq \sqrt[3]{\eta^*}\varepsilon_0,\ |\eta-\eta^*|\leq\varepsilon_0 \quad (3.10.2)$$

with probability at least $1-CO(1/n)$. Assume the step size μ satisfies $0<\mu<\mu_0$ for some small absolute constant μ_0 and $s\leq d\leq Cs$. Then $\{\widetilde{\boldsymbol{\beta}}_j^{(t+1)}\}$ can be upper bounded as

$$\begin{aligned}&\max_{j=1,2,3}\Big\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_j^{(t+1)}-\sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\Big\|_2\\ \leq\ & (1-\frac{\mu}{12})\max_{j=1,2,3}\Big\|\sqrt[3]{\eta}\boldsymbol{\beta}_j^{(t)}-\sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\Big\|_2+\mu\frac{3\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{3s\log p}{n}},\end{aligned}$$

with probability at least $1-12s/n$.

Proof. We focus on $j=1$ first. To simplify the notation, we drop the superscript of iteration index t, and denote iteration index $t+1$ by $+$. Moreover, denote $\bar{\boldsymbol{\beta}}_j=\sqrt[3]{\eta}\boldsymbol{\beta}_j$, $\bar{\boldsymbol{\beta}}_j^+=\sqrt[3]{\eta}\boldsymbol{\beta}_j$, $\bar{\boldsymbol{\beta}}_j^*=\sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*$ for $j=1,2,3$. Then, the gradient function is rewritten as

$$\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)=\sqrt[3]{\eta}\frac{2}{n}\sum_{i=1}^{n}\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\Big)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\boldsymbol{u}_i.$$

According to the definition of thresholded function, $\widetilde{\boldsymbol{\beta}}_1^+$ can be explicitly written by

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_1^+ &= \varphi_{\frac{\mu}{\phi}h(\bar{\boldsymbol{\beta}}_1)}\Big(\boldsymbol{\beta}_1-\frac{\mu}{\phi}\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)_F\Big)\\ &= \boldsymbol{\beta}_1-\frac{\mu}{\phi}\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)_F+\frac{\mu}{\phi}h(\bar{\boldsymbol{\beta}}_1)\boldsymbol{\gamma},\end{aligned}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^p, \text{supp}(\boldsymbol{\gamma}) \subset F$ and $\|\boldsymbol{\gamma}\|_\infty \leq 1$. Then the oracle estimation error $\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\|_2$ can be decomposed by the gradient update effect and the thresholded effect,

$$\begin{aligned} &\left\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\right\|_2 \\ =&\underbrace{\left\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\right\|_2}_{\text{gradient update effect}} + \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\left|h(\bar{\boldsymbol{\beta}}_1)\right|\sqrt{3s}}_{\text{thresholded effect}}. \end{aligned} \tag{3.10.3}$$

By using the tri-convex structure of $\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)$, we borrow the analysis tool for vanilla gradient descent (95) given sufficient good initial. Following this proof strategy, we decompose the gradient update effect in (3.10.3) by three parts,

$$\begin{aligned} \left\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\right\|_2 \quad \leq \quad & \underbrace{\left\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)_F\right\|_2}_{I_1} \\ + \quad & \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\left\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)_F - \nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\right\|_2}_{I_2} \\ + \quad & \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\left\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F - \nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\right\|_2}_{I_3} \\ + \quad & \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\left|h(\bar{\boldsymbol{\beta}}_1)\right|\sqrt{3s}}_{I_4}, \end{aligned}$$

where $\nabla_1\widetilde{\mathcal{L}}$ is the noiseless gradient as we defined in (3.9.30). We will bound I_1, I_2, I_3, I_4 successively in the following four subsections. For simplicity, during the following proof, we drop the index subscript F as we did in Section 3.9.12. And $\phi = \sum_{i=1}^n y_i^2$ approximates η^{*2} up to constant due to Lemma 15.

Bounding I_1 In this section, let us denote

$$\sqrt[3]{\eta}\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)/\phi = f(\bar{\boldsymbol{\beta}}_1), \ \sqrt[3]{\eta}\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)/\phi = \nabla f(\bar{\boldsymbol{\beta}}_1),$$

where $\text{supp}(\nabla f(\bar{\boldsymbol{\beta}}_1)) = F$. When $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are fixed, the update can be treated as a vanilla gradient descent update. The following proof follows three steps. The first two steps show that $f(\bar{\boldsymbol{\beta}}_1)$ is Lipshitz differentiable and strongly convex on the constraint set F, and the last step utilizes the classical convex gradient analysis.

Step One: Verify $f(\bar{\boldsymbol{\beta}}_1)$ is L-Lipschitz differentiable. For any $\bar{\boldsymbol{\beta}}_1^{(1)}$ and $\bar{\boldsymbol{\beta}}_1^{(2)}$ whose support belong to F,

$$\nabla f(\bar{\boldsymbol{\beta}}_1^{(1)}) - \nabla f(\bar{\boldsymbol{\beta}}_1^{(2)}) = \frac{(\sqrt[3]{\eta})^2}{\phi}\frac{2}{n}\sum_{i=1}^{n}\Big(\boldsymbol{u}_i^\top(\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)})(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2\Big)\boldsymbol{u}_i.$$

Then, there exist $\pi \in \mathbb{S}^{s-1}$ such that

$$\begin{aligned}&\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1^{(1)}) - \nabla f(\bar{\boldsymbol{\beta}}_1^{(2)})\Big\|_2\\ =\ & \frac{(\sqrt[3]{\eta})^2}{\phi}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{u}_i^\top(\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)})(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2\Big)\boldsymbol{u}_i^\top\pi\Big|.\end{aligned}$$

Applying Lemma 2 with multiplying $(\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)}) \circ \bar{\boldsymbol{\beta}}_2^* \circ \bar{\boldsymbol{\beta}}_3^*$, it shows

$$\begin{aligned}&\Big|\sum_{i=1}^{n}\Big[(\boldsymbol{u}_i^\top(\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)})(\boldsymbol{u}_i^\top\pi)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2)\Big]\Big|\\ \le\ & \Big(1+\delta_{n,p,s}\Big)\Big\|\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)}\Big\|_2\eta^{*\frac{4}{3}},\end{aligned}$$

with probability at least $1 - 10/n^3$, where $\delta_{n,p,s}$ is defined in (3.3.7). Under Condition (5) with some constant adjustments, we obtain

$$\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1^{(1)}) - \nabla f(\bar{\boldsymbol{\beta}}_1^{(2)})\Big\|_2 \le \frac{57}{16}\Big\|\bar{\boldsymbol{\beta}}_1^{(1)} - \bar{\boldsymbol{\beta}}_1^{(2)}\Big\|_2. \tag{3.10.4}$$

with probability at least $1 - 10/n^3$. Therefore, $f(\bar{\boldsymbol{\beta}}_1)$ is Lipschitz differentiable with Lipschitz constant $L = \frac{57}{8}$.

Step Two: Verify $f(\bar{\boldsymbol{\beta}}_1)$ is α-strongly convex. It is equivalent to prove that $\nabla^2 f(\bar{\boldsymbol{\beta}}_1) \succeq m\mathbb{I}_p$. Based on the inequality (3.3.19) in (81), it shows that

$$\lambda_{\min}\big(\nabla^2(f(\bar{\boldsymbol{\beta}}_1))\big) \ge \lambda_{\min}\big(\mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\big) - \lambda_{\max}\big(\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\big). \tag{3.10.5}$$

The lower bound of $\lambda_{\min}(\nabla^2(f(\bar{\boldsymbol{\beta}}_1)))$ breaks into two parts: an lower bound for $\lambda_{\min}(\mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1)))$, and an upper bound for $\lambda_{\max}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))$. The Hessian matrix of $f(\bar{\boldsymbol{\beta}}_1)$ is given by

$$\nabla^2 f(\bar{\boldsymbol{\beta}}_1) = \frac{(\sqrt[3]{\eta})^2}{\phi}\frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{v}_1^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2\boldsymbol{u}_i\boldsymbol{u}_i^\top.$$

Since $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i$ are independent with each other, we have $\mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1)) = 2\mathbb{I}$, which implies $\lambda_{\min}\big(\mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\big) \geq 2$.

On the other hand,

$$\begin{aligned}
&\lambda_{\max}\Big(\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\Big) = \Big\|\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\Big\|_2 \\
&\leq \boldsymbol{a}^\top\Big(\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\Big)\boldsymbol{b} = \frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{v}_i^\top \bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top \bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{u}_i^\top \boldsymbol{a})(\boldsymbol{u}_i^\top \boldsymbol{b}) \\
&-\mathbb{E}\Big(\sum_{i=1}^{n}(\boldsymbol{v}_i^\top \bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top \bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{u}_i^\top \boldsymbol{a})(\boldsymbol{u}_i^\top \boldsymbol{b})\Big)\eta^{*-\frac{4}{3}}.
\end{aligned}$$

where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{S}^{s-1}$. Equipped with Lemma 2, it yields that with probability at least $1 - 10/n^3$,

$$\lambda_{\max}\Big(\nabla^2 f(\bar{\boldsymbol{\beta}}_1) - \mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1))\Big) \leq 2\delta_{n,s,p}.$$

Together with the lower bound of $\lambda_{\min}(\mathbb{E}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1)))$, we have

$$\lambda_{\min}(\nabla^2 f(\bar{\boldsymbol{\beta}}_1)) \geq 2 - 2\delta_{n,p,s},$$

Under Condition 5, the minimum eigenvalue of Hessian matrix $\nabla^2 f(\bar{\boldsymbol{\beta}}_1)$ is lower bounded by $\frac{19}{10}$ with probability at least $1 - 10/n^3$. This guarantees that $f(\bar{\boldsymbol{\beta}}_1)$ is strongly-convex with $\alpha = \frac{19}{10}$.

Step Three: Combining the Lipschitz condition, strongly-convexity and Lemma 3.11 in (95), it shows that

$$\begin{aligned}
&\big(\nabla f(\bar{\boldsymbol{\beta}}_1) - \nabla f(\bar{\boldsymbol{\beta}}_1^*)^\top\big)\big(\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}^*\big) \\
\geq\ & \frac{\alpha L}{\alpha + L}\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*\Big\|_2^2 + \frac{1}{\alpha + L}\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1) - \nabla f(\bar{\boldsymbol{\beta}}_1^*)\Big\|_2^2.
\end{aligned}$$

Since the gradient vanishes at the optimal point, the above inequality times 2μ simplifies to

$$-2\mu\nabla f(\bar{\boldsymbol{\beta}}_1)^\top(\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*) \leq -\frac{2\mu\alpha L}{\alpha + L}\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*\Big\|_2^2 - \frac{2\mu}{\alpha + L}\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1)\Big\|_2^2. \tag{3.10.6}$$

Now it's sufficient to bound $\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\nabla f(\bar{\boldsymbol{\beta}}_1)\|_2$ as follows

$$\begin{aligned}
& \Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\nabla f(\bar{\boldsymbol{\beta}}_1)\Big\|_2^2 \\
= \; & \Big\|\bar{\boldsymbol{\beta}}_1^t - \bar{\boldsymbol{\beta}}_1^*\Big\|_2^2 + \mu^2\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1)\Big\|_2^2 - 2\mu\nabla f(\bar{\boldsymbol{\beta}}_1)^\top(\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}^*) \\
\leq \; & \Big(1 - 2\mu\frac{\alpha L}{\alpha + L}\Big)\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*\Big\|_2^2 + \mu\big(\mu - \frac{2}{\alpha + L}\big)\Big\|\nabla f(\bar{\boldsymbol{\beta}}_1)\Big\|_2^2.
\end{aligned}$$

where L, α are Lipschitz constant and strongly convexity parameter, respectively. If $\mu < \frac{80}{361}$, the last term can be neglected and we obtain the desired upper bound,

$$\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\widetilde{\mathcal{L}}\big(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*\big)\Big\|_2 \leq \Big(1 - 3\mu\Big)\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*\Big\|_2, \tag{3.10.7}$$

with probability $1 - 20/n^3$. This ends the proof. ■

Bounding I_2 For simplicity, we write $\boldsymbol{z}_1 = \bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^*$, $\boldsymbol{z}_2 = \bar{\boldsymbol{\beta}}_2 - \bar{\boldsymbol{\beta}}_2^*$, $\boldsymbol{z}_3 = \bar{\boldsymbol{\beta}}_3 - \bar{\boldsymbol{\beta}}_2^*$. By the definition of noiseless gradient, it suffices to decompose I_2 by

$$\begin{aligned}
& \eta^{-\frac{1}{3}}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*) - \nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)\Big\|_2 \\
\leq \; & \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{v}_i^\top\boldsymbol{z}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)(\boldsymbol{w}_i^\top\boldsymbol{z}_3)\boldsymbol{u}_i\Big\|_2 \\
& + \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{v}_i^\top\boldsymbol{z}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)\boldsymbol{u}_i\Big\|_2 \\
& + \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)(\boldsymbol{w}_i^\top\boldsymbol{z}_3)\boldsymbol{u}_i\Big\|_2 \\
& + \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{v}_i^\top\boldsymbol{z}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)(\boldsymbol{w}_i^\top\boldsymbol{z}_3)\boldsymbol{u}_i\Big\|_2 \\
& + \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{v}_i^\top\boldsymbol{z}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2\boldsymbol{u}_i\Big\|_2 \\
& + \Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)(\boldsymbol{w}_i^\top\boldsymbol{z}_3)\boldsymbol{u}_i\Big\|_2.
\end{aligned}$$

Repeatedly using Lemma 2, we obtain

$$\begin{aligned}
&\eta^{-\frac{1}{3}}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2^*,\bar{\boldsymbol{\beta}}_3^*)-\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)\Big\|_2\\
&\leq\Big(1+\delta_{n,p,s}\Big)\Big[(1+\varepsilon_0)^3\varepsilon_0+(1+\varepsilon_0)^3+(1+\varepsilon_0)^3+\varepsilon_0^2+2\varepsilon_0\Big]\eta^{*\frac{4}{3}}\max_j\|\boldsymbol{z}_j\|_2\\
&\leq\frac{5}{2}\Big(1+\delta_{n,p,s}\Big)\eta^{*\frac{4}{3}}\max_j\|\boldsymbol{z}_j\|_2,
\end{aligned}$$

for sufficiently small ε_0 with probability at least $1-60/n^3$. Under Condition 5, it suffices to get

$$\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)-\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2^*,\bar{\boldsymbol{\beta}}_3^*)\Big\|_2\leq\frac{8}{3}\max_{j=1,2,3}\Big\|\bar{\boldsymbol{\beta}}_j-\bar{\boldsymbol{\beta}}_j^*\Big\|_2,\tag{3.10.8}$$

with probability at least $1-6/n$. ■

Bounding I_3 I_3 quantifies the statistical error. By the definition of noiseless gradient and noisy gradient, we have

$$\begin{aligned}
&\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)-\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)\Big\|_2\\
=\;&\frac{(\sqrt[3]{\eta})^2}{\phi}\Big\|\frac{2}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\boldsymbol{u}_i\Big\|_2.
\end{aligned}$$

The proof of this part essentially coincides with the proof for symmetric tensor estimation. Combining Lemmas 1 and 23, we have

$$\Big|\frac{2}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)u_{ij}\Big|\leq C(1+\varepsilon_0)^2\eta^{*\frac{2}{3}}\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}},$$

with probability at least $1-4/n$. Applying union bound over $3s$ coordinates, it suffices to get

$$\mathbb{P}\Big(\max_{j\in[3s]}\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)u_{ij}\Big|\geq C(1+\varepsilon_0)^2\eta^{*-\frac{2}{3}}\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}\Big)\leq\frac{12s}{n}.$$

Therefore, we reach

$$\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)-\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)\Big\|_2\leq 2C\eta^{*-\frac{2}{3}}\sigma\sqrt{\frac{3s(\log n)^3}{n}},\tag{3.10.9}$$

with probability at least $1-12s/n$. ■

Bounding I_4 According to the definition of thresholding level $h(\bar{\boldsymbol{\beta}}_1)$ in (3.10.1), we can bound the square as follows,

$$\frac{(\sqrt[3]{\eta})^2}{\phi^2}h^2(\bar{\boldsymbol{\beta}}_1) = \frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^{n}\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3) - (\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*) - \epsilon_i\Big)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2$$

Based on the basic inequality $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned}&\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3) - (\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*) - \epsilon_i\Big)^2\\ \le\ & 2\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3) - (\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)\Big)^2 + 2\epsilon_i^2.\end{aligned}$$

Denote I_1 and I_2 corresponding to optimization error and statistical error,

$$\begin{aligned}I_1 &= \frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^{n}\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\\ &\quad -(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)\Big)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ I_2 &= \frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^{n}\epsilon_i^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2.\end{aligned}$$

Next, I_1 is decomposed by some high-order polynomials as follows

$$\begin{aligned}I_1 =& \frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\Big(\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\ &+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\Big).\end{aligned} \tag{3.10.10}$$

Each term contains the product of Gaussian random vectors form up to power ten. For the first term, by using Lemma 1,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i^\top \boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top \boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top \boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top \bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top \bar{\boldsymbol{\beta}}_3)^2 \\ \leq \; & (1+\varepsilon_0)^4\varepsilon_0^4\Big(1+C\frac{(\log n)^5}{\sqrt{n}}\Big)\eta^{*\frac{8}{3}}\max_{j=1,2,3}\|\boldsymbol{z}_j\|_2^2, \end{aligned}$$

with probability at least $1-1/n$. Similar bounds holds for other terms. As long as $n \geq C\log^{10} n$, we have with probability at least $1-7/n$,

$$I_1 \leq \frac{7\log p}{n}\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_j^*\right\|_2^2. \tag{3.10.11}$$

Now we turn to bound I_2. For fixed $\{\epsilon_i\}$, we have,

$$\begin{aligned} & \Big|\sum_{i=1}^{n}\epsilon_i^2(\boldsymbol{v}_i^\top \bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top \bar{\boldsymbol{\beta}}_3)^2 - \sum_{i=1}^{n}\epsilon_i^2\|\bar{\boldsymbol{\beta}}_2\|_2^2\|\bar{\boldsymbol{\beta}}_3\|_2^2\Big| \\ \leq \; & C(\log n)^2\|\boldsymbol{\epsilon}^2\|_2\|\bar{\boldsymbol{\beta}}_2\|_2^2\|\bar{\boldsymbol{\beta}}_3\|_2^2, \end{aligned}$$

with probability at least $1-n^{-1}$. Combining with Lemma 23,

$$I_2 \leq 4\sigma^2\eta^{*\frac{4}{3}}\frac{\log p}{n}. \tag{3.10.12}$$

Putting (3.10.11) and (3.10.12) together, the thresholded effect can be bound by

$$\frac{\sqrt[3]{\eta}}{\phi}|h(\boldsymbol{\beta}_1)| \leq \sqrt{\frac{7\log np}{n}}\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_j^*\right\|_2 + \frac{2\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{\log np}{n}}, \tag{3.10.13}$$

with probability at least $1-8/n$, provided $n \gtrsim (\log n)^{10}$. ■

Summary Putting the upper bounds (3.10.7), (3.10.8) and (3.10.13) together, we obtain that if step size μ satisfies $0 < \mu < \mu_0$ for some small μ_0,

$$\left\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\right\|_2 \leq (1-\frac{\mu}{12})\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_j^*\right\|_2 + \mu\frac{3\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{3s\log p}{n}},$$

with probability at least $1-12s/n$. This finishes our proof. ■

3.11 Matrix Form Gradient and Stochastic Gradient descent

3.11.1 Matrix Formulation of Gradient

In this section, we provide detail derivations for (3.2.7) and (3.5.5).

Lemma 3.11.1. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \in \mathbb{R}^{K\times 1}, \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{p\times n}$ and $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times K}$. The gradient of symmetric tensor estimation empirical risk function (3.2.5) can be written in a matrix form as follows

$$\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \frac{6}{n}[((\boldsymbol{B}^\top \boldsymbol{X})^\top)^3\boldsymbol{\eta} - \boldsymbol{y}]^\top[(((\boldsymbol{B}^\top \boldsymbol{X})^\top)^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{X}]^\top.$$

Proof. First let's have a look at the gradient for k-th component,

$$\nabla\mathcal{L}_k(\boldsymbol{\beta}_k) = \frac{6}{n}(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3 - y_i)\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)\boldsymbol{x}_i \in \mathbb{R}^{p\times 1}, \quad \text{for } k = 1, \ldots, K.$$

Correspondingly, each part can be written as a matrix form,

$$((\underbrace{\boldsymbol{B}^\top\boldsymbol{X}}_{K\times n})^\top)^3\boldsymbol{\eta} - \boldsymbol{y} \in \mathbb{R}^{n\times 1}$$
$$(((\boldsymbol{B}^\top \boldsymbol{X})^\top)^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{X} \in \mathbb{R}^{pK\times n}.$$

This implies that $[((\boldsymbol{B}^\top \boldsymbol{X})^\top)^3\eta - \boldsymbol{y}]^\top[(((\boldsymbol{B}^\top \boldsymbol{X})^\top)^2 \odot \eta^\top)^\top \odot \boldsymbol{X}]^\top \in \mathbb{R}^{1\times pK}$. Note that $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = (\nabla\mathcal{L}_1(\boldsymbol{\beta}_1)^\top, \ldots, \nabla\mathcal{L}_K(\boldsymbol{\beta}_K)^\top) \in \mathbb{R}^{1\times pK}$. The conclusion can be easily derived. ■

Lemma 17 . Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \in \mathbb{R}^{K\times 1}, \boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \in \mathbb{R}^{p_1\times n}, \boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{p_2\times n}, \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) \in \mathbb{R}^{p_3\times n}$ and $\boldsymbol{B}_1 = (\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1K}) \in \mathbb{R}^{p_1\times K}, \boldsymbol{B}_2 = (\boldsymbol{\beta}_{21}, \ldots, \boldsymbol{\beta}_{2K}) \in \mathbb{R}^{p_2\times K}, \boldsymbol{B}_3 = (\boldsymbol{\beta}_{31}, \ldots, \boldsymbol{\beta}_{3K}) \in \mathbb{R}^{p_3\times K}$. The gradient of non-symmetric tensor estimation empirical risk function (3.5.3) can be written in a matrix form as follows

$$\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{\eta}) = \boldsymbol{D}^\top(\boldsymbol{C}_1^\top \odot \boldsymbol{U})^\top,$$

where $\boldsymbol{D} = (\boldsymbol{B}_1^\top\boldsymbol{U})^\top * (\boldsymbol{B}_2^\top\boldsymbol{V})^\top * (\boldsymbol{B}_3^\top\boldsymbol{W})^\top\boldsymbol{\eta} - \boldsymbol{y}$ and $\boldsymbol{C}_1 = (\boldsymbol{B}_2^\top\boldsymbol{V})^\top * (\boldsymbol{B}_3^\top\boldsymbol{W})^\top \odot \boldsymbol{\eta}^\top$.

Proof. Recall that $\{*, \odot\}$ represent Hadamard product and Khatri-Rao product respectively. Then the dimensionality of $\boldsymbol{D}, \boldsymbol{C}_1, \boldsymbol{C}_1 \odot \boldsymbol{U}$ can be calculated as follows

$$\begin{aligned}
\boldsymbol{D} &= \underbrace{(\boldsymbol{B}_1^\top \boldsymbol{U})^\top}_{n\times K} * \underbrace{(\boldsymbol{B}_2^\top \boldsymbol{V})^\top}_{n\times K} * \underbrace{(\boldsymbol{B}_3^\top \boldsymbol{W})^\top}_{n\times K} \boldsymbol{\eta} - \boldsymbol{y} \in \mathbb{R}^{n\times 1}, \\
\boldsymbol{C}_1 &= (\boldsymbol{B}_2^\top \boldsymbol{V})^\top * (\boldsymbol{B}_3^\top \boldsymbol{W})^\top \odot \boldsymbol{\eta}^\top \in \mathbb{R}^{n\times K}, \boldsymbol{C}_1^\top \odot \boldsymbol{U} \in \mathbb{R}^{Kp_1\times n}.
\end{aligned}$$

Therefore,

$$\nabla_{\boldsymbol{B}_1} \mathcal{L}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{\eta}) = \boldsymbol{D}^\top (\boldsymbol{C}_1^\top \odot \boldsymbol{U})^\top = (\nabla_1 \mathcal{L}(\boldsymbol{\beta}_1)^\top, \ldots, \nabla_K \mathcal{L}(\boldsymbol{\beta}_K)^\top).$$

■

3.11.2 Stochastic Gradient descent

Stochastic thresholded gradient descent is a stochastic approximation of the gradient descent optimization method. Note that the empirical risk function (3.2.5) that can be written as a sum of differentiable functions. Followed by (3.2.7), the gradient of (3.2.5) evaluated at i-th sketching $\{y_i, \boldsymbol{x}_i\}$ can be written as

$$\nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}, \boldsymbol{\eta}) = [((\boldsymbol{B}^\top \boldsymbol{x}_i)^\top)^3 \boldsymbol{\eta} - y_i][(((\boldsymbol{B}^\top \boldsymbol{x}_i)^\top)^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{x}_i]^\top \in \mathbb{R}^{1\times pK},$$

Thus, the overall gradient $\nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}, \boldsymbol{\eta})$ defined in (3.2.7) can be expressed as a summand of $\nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}, \boldsymbol{\eta})$,

$$\nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}, \boldsymbol{\eta}).$$

The thresholded step remains the same as Step 3 in Algorithm1. Then the symmetric update of stochastic thresholded gradient descent within one iteration is summarized by

$$\mathrm{vec}(\boldsymbol{B}^{(t+1)}) = \varphi_{\frac{\mu_{SGD}}{\phi} \boldsymbol{h}(\boldsymbol{B}^{(t)})} \Big(\mathrm{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu_{SGD}}{\phi} \nabla_{\boldsymbol{B}} \mathcal{L}_i(\boldsymbol{B}^{(t)}) \Big).$$

3.12 Technical Lemmas

Lemma 18 . Suppose $\boldsymbol{x} \in \mathbb{R}^p$ is a standard Gaussian random vector. For any non-random vector $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^p$, we have the following tensor expectation calculation,

$$\begin{aligned}
&\mathbb{E}\Big((\boldsymbol{a}^\top \boldsymbol{x})(\boldsymbol{b}^\top \boldsymbol{x})(\boldsymbol{c}^\top \boldsymbol{x})\boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x}\Big) \\
&= \Big(\boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c} + \boldsymbol{a} \circ \boldsymbol{c} \circ \boldsymbol{b} + \boldsymbol{b} \circ \boldsymbol{a} \circ \boldsymbol{c} + \boldsymbol{b} \circ \boldsymbol{c} \circ \boldsymbol{a} + \boldsymbol{c} \circ \boldsymbol{b} \circ \boldsymbol{a} + \boldsymbol{c} \circ \boldsymbol{a} \circ \boldsymbol{b}\Big) \\
&+ 3 \sum_{m=1}^{p} \Big(\boldsymbol{a} \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m (\boldsymbol{b}^\top \boldsymbol{c}) + \boldsymbol{e}_m \circ \boldsymbol{b} \circ \boldsymbol{e}_m (\boldsymbol{a}^\top \boldsymbol{c}) + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ \boldsymbol{c} (\boldsymbol{a}^\top \boldsymbol{b})\Big),
\end{aligned} \tag{3.12.1}$$

where $\boldsymbol{e}_m$ is a canonical vector in $\mathbb{R}^p$.

Proof. Recall that for a standard Gaussian random variable x, its odd moments are zero and even moments are $\mathbb{E}(x^6) = 15, \mathbb{E}(x^4) = 4$. Expanding the LHS of (3.12.1) and comparing LHS and RHS, we will reach the conclusion. Details are omitted here. ■

Lemma 19 . Suppose $\boldsymbol{u} \in \mathbb{R}^{p_1}, \boldsymbol{v} \in \mathbb{R}^{p_2}, \boldsymbol{w} \in \mathbb{R}^{p_3}$ are independent standard Gaussian random vectors. For any non-random vector $\boldsymbol{a} \in \mathbb{R}^{p_1}, \boldsymbol{b} \in \mathbb{R}^{p_2}, \boldsymbol{c} \in \mathbb{R}^{p_3}$, we have the following tensor expectation calculation

$$\mathbb{E}\Big((\boldsymbol{a}^\top \boldsymbol{u})(\boldsymbol{b}^\top \boldsymbol{v})(\boldsymbol{c}^\top \boldsymbol{w})\boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\Big) = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c}. \tag{3.12.2}$$

Proof. Due to the independence among $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$, the conclusion is easy to obtain by using the moment of standard Gaussian random variable. ■

Note that in the left side of (3.12.1), it involves an expectation of rank-one tensor. When multiplying any non-random rank-one tensor with same dimensionality, i.e. $\boldsymbol{a}_1 \circ \boldsymbol{b}_1 \circ \boldsymbol{c}_1$, on both sides, it will facilitate us to calculate the expectation of product of Gaussian vectors, see next Lemma for details.

Lemma 3.12.1. Suppose $\boldsymbol{x} \in \mathbb{R}^p$ is a standard Gaussian random vector. For any non-random vector $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^p$, we have the following expectation calculation

$$
\begin{aligned}
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^6 = 15\|\boldsymbol{a}\|_2^6, \\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^5(\boldsymbol{x}^\top \boldsymbol{b}) = 15\|\boldsymbol{a}\|_2^4(\boldsymbol{a}^\top \boldsymbol{b}), \\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^4(\boldsymbol{x}^\top \boldsymbol{b})^2 = 12\|\boldsymbol{a}\|_2^2(\boldsymbol{a}^\top \boldsymbol{b})^2 + 3\|\boldsymbol{a}\|_2^4\|\boldsymbol{b}\|_2^2, \\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^3 = 6(\boldsymbol{a}^\top \boldsymbol{b})^3 + 9(\boldsymbol{a}^\top \boldsymbol{b})\|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2, \\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^2(\boldsymbol{x}^\top \boldsymbol{c}) = 6(\boldsymbol{a}^\top \boldsymbol{b})^2(\boldsymbol{a}^\top \boldsymbol{c}) + 6(\boldsymbol{a}^\top \boldsymbol{b})(\boldsymbol{b}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{a}) \\
&+3(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{b}^\top \boldsymbol{b})(\boldsymbol{a}^\top \boldsymbol{a}), \\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^2(\boldsymbol{x}^\top \boldsymbol{b})(\boldsymbol{x}^\top \boldsymbol{c})^2(\boldsymbol{x}^\top \boldsymbol{d}) = 2(\boldsymbol{a}^\top \boldsymbol{c})^2(\boldsymbol{b}^\top \boldsymbol{d}) + 4(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{b}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{d}) \\
&+6(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{b})(\boldsymbol{c}^\top \boldsymbol{d}) + 3(\boldsymbol{c}^\top \boldsymbol{x})(\boldsymbol{b}^\top \boldsymbol{d})(\boldsymbol{a}^\top \boldsymbol{a}).
\end{aligned}
$$

Proof. Note that $\mathbb{E}((\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^3) = \mathbb{E}((\boldsymbol{x}^\top \boldsymbol{a})^3\langle \boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}, \boldsymbol{b}\circ\boldsymbol{b}\circ\boldsymbol{b}\rangle)$. Then we can apply the general result in Lemma 18. Comparing both sides, we will obtain the conclusion. Others part follows the similar strategy. ■

Next lemma provides a probabilistic concentration bound for non-symmetric rank-one tensor under tensor spectral norm.

Lemma 20 . Suppose $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_n^\top)^\top, \boldsymbol{Y} = (\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_n^\top)^\top, \boldsymbol{Z} = (\boldsymbol{z}_1^\top, \cdots, \boldsymbol{z}_n^\top)^\top$ are three $n \times p$ random matrices. The ψ_2-norm of each entry is bounded, s.t. $\|X_{ij}\|_{\psi_2} = K_x, \|Y_{ij}\|_{\psi_2} = K_y, \|Z_{ij}\|_{\psi_2} = K_z$. We assume the row of $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ are independent. There exists an absolute constant C such that,

$$
\begin{aligned}
&\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \Big[\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i - \mathbb{E}(\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i)\Big]\Big\|_s \geq CK_xK_yK_z\delta_{n,p,s}\Big) \leq p^{-1}. \\
&\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \Big[\boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i)\Big]\Big\|_s \geq CK_x^3\delta_{n,p,s}\Big) \leq p^{-1}.
\end{aligned}
$$

Here, $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (3.1.3) and $\delta_{n,p,s} = \sqrt{s\log(ep/s)/n} + \sqrt{s^3\log(ep/s)^3/n^2}$.

Proof. Bounding spectral norm always relies on the construction of the ϵ-net. Since we will bound a sparse tensor spectral norm, our strategy is to discrete the sparse set and construct the ϵ-net on each one. Let us define a sparse set $\mathcal{B}_0 = \{\boldsymbol{x} \in \mathbb{R}^p, \|\boldsymbol{x}\|_2 = 1, \|\boldsymbol{x}\|_0 \leq s\}$. And let $\mathcal{B}_{0,s}$ be the s-dimensional set defined by $\mathcal{B}_{0,s} = \{\boldsymbol{x} \in \mathbb{R}^s, \|\boldsymbol{x}\|_2 = 1\}$. Note that $\mathcal{B}_0$ is corresponding to s-sparse unit vector set which can be expressed as a union of subsets of dimension s by expanding some zeros, namely $\mathcal{B}_0 = \cup\, \mathcal{B}_{0,s}$. There should be at most $\binom{p}{s} \leq (\frac{ep}{s})^s$ such set $\mathcal{B}_{0,s}$.

Recalling the definition of sparse tensor spectral norm in (3.1.3), we have

$$\begin{aligned} A &= \Big\| \frac{1}{n} \sum_{i=1}^{n} \Big[\boldsymbol{x}_i \circ \boldsymbol{y}_i \circ \boldsymbol{z}_i - \mathbb{E}(\boldsymbol{x}_i \circ \boldsymbol{y}_i \circ \boldsymbol{z}_i) \Big] \Big\|_s \\ &= \sup_{\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3 \in \mathcal{B}_0} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big[\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle - \mathbb{E}(\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle) \Big] \Big|. \end{aligned}$$

Instead of constructing the ϵ-net on $\mathcal{B}_0$, we will construct an ϵ-net for each of subsets $\mathcal{B}_{0,s}$. Define $\mathcal{N}_{\mathcal{B}_{0,s}}$ as the 1/2-set of $\mathcal{B}_{0,s}$. From Lemma 3.18 in (102), the cardinality of $\mathcal{N}_{0,s}$ is bounded by 5^s. By Lemma 21, we obtain

$$\begin{aligned} &\sup_{\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3 \in \mathcal{B}_{0,s}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big[\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle - \mathbb{E}(\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle) \Big] \Big| \\ \leq & 2^3 \sup_{\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3 \in \mathcal{N}_{\mathcal{B}_{0,s}}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Big[\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle - \mathbb{E}(\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle \langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle \langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle) \Big] \Big|. \end{aligned} \tag{3.12.3}$$

By rotation invariance of sub-Gaussian random variable, $\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1 \rangle$, $\langle \boldsymbol{y}_i, \boldsymbol{\chi}_2 \rangle$, $\langle \boldsymbol{z}_i, \boldsymbol{\chi}_3 \rangle$ are still sub-Gaussian random variables with ψ_2-norm bounded by K_x, K_y, K_z, respectively. Applying Lemma 1 and union bound over $\mathcal{N}_{\mathcal{B}_{0,s}}$, the right hand side of (3.12.3) can be bounded by

$$\mathbb{P}\Big(\text{RHS} \geq 8 K_x K_y K_z C \Big(\sqrt{\frac{\log \delta^{-1}}{n}} + \sqrt{\frac{(\log \delta^{-1})^3}{n^2}} \Big) \Big) \leq (5^s)^3 \delta,$$

for any $0 < \delta < 1$.

Lastly, taking the union bound over all possible subsets $\mathcal{B}_{0,s}$ yields that

$$\begin{aligned} & \mathbb{P}\Big(A \geq 8 K_x K_y K_z C \Big(\sqrt{\frac{\log \delta^{-1}}{n}} + \sqrt{\frac{(\log \delta^{-1})^3}{n^2}} \Big) \Big) \\ \leq & \ (\frac{ep}{s})^s (5^s)^3 \delta = (\frac{125ep}{s})^s \delta. \end{aligned}$$

Letting $p^{-1} = (\frac{125ep}{s})^s\delta$, we obtain with probability at least $1 - 1/p$

$$A \leq CK_xK_yK_z\Big(\sqrt{\frac{s\log(p/s)}{n}} + \sqrt{\frac{s^3\log^3(p/s)}{n^2}}\Big),$$

with some adjustments on constant C. The proof for symmetric case is similar to non-symmetric case so we omit here. ■

Lemma 21 (Tensor Covering Number(Lemma 4 in (103))). Let $\mathbb{N}$ be an ϵ-net for a set $\boldsymbol{B}$ associated with a norm $\|\cdot\|$. Then, the spectral norm of a d-mode tensor $\mathcal{A}$ is bounded by

$$\begin{aligned}&\sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{d-1}\in\boldsymbol{B}} \|\mathcal{A}\times_1\boldsymbol{x}_1\ldots\times_{d-1}\boldsymbol{x}_{d-1}\|_2\\ &\leq \Big(\frac{1}{1-\varepsilon}\Big)^{d-1}\sup_{\boldsymbol{x}_1\cdots\boldsymbol{x}_{d-1}\in\mathbb{N}}\|\mathcal{A}\times_1\boldsymbol{x}_1\cdots\times_{d-1}\boldsymbol{x}_{d-1}\|_2.\end{aligned}$$

This immediately implies that the spectral norm of a d-mode tensor $\mathcal{A}$ is bounded by

$$\|\mathcal{A}\|_2 \leq (\frac{1}{1-\epsilon})^{d-1}\sup_{\boldsymbol{x}_1\ldots\boldsymbol{x}_{d-1}\in\mathcal{N}}\|\mathcal{A}\times_1\boldsymbol{x}_1\cdots\times_{d-1}\boldsymbol{x}_{d-1}\|_2,$$

where $\mathbb{N}$ is the ϵ-net for the unit sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$.

Lemma 22 (Sub-Gaussianess of the Product of Random Variables). Suppose X_1 is a bounded random variable with $|X_1| \leq K_1$ almost surely for some K_1 and X_2 is a sub-Gaussian random variable with Orlicz norm $\|X_2\|_{\psi_2}K_2$. Then X_1X_2 is still a sub-Gaussian random variable with Orlicz norm $\|X_1X_2\|_{\psi_2} = K_1K_2$.

Proof: Following the definition of sub-Gaussian random variable, we have

$$\mathbb{P}\Big(\big|X_1X_2\big| > t\Big) = \mathbb{P}\Big(\big|X_2\big| > \frac{t}{\big|X_1\big|}\Big) \leq \mathbb{P}\Big(\big|X_2\big| > \frac{t}{\big|K_1\big|}\Big) \leq \exp\Big(1 - t^2/K_1^2K_2^2\Big),$$

holds for all $t \geq 0$. This ends the proof. ■

Lemma 23 (Tail Probability for the Sum of Sub-exponential Random Variables (Lemma A.7 in (89))). Suppose $\epsilon_1, \ldots, \epsilon_n$ are independent centered sub-exponential random variables with

$$\sigma := \max_{1 \le i \le n} \|\epsilon_i\|_{\psi_1}.$$

Then with probability at least $1 - 3/n$, we have

$$\Big|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i\Big| \le C_0\sigma\sqrt{\frac{\log n}{n}}, \ \big\|\boldsymbol{\epsilon}\big\|_\infty \le C_0\sigma \log n,$$
$$\Big|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2\Big| \le C_0\sigma^2, \Big|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^4\Big| \le C_0\sigma^4,$$

for some constant C_0.

Lemma 24 (Tail Probability for the Sum of Weibull Distributions (Lemma 3.6 in (57))). Let $\alpha \in [1, 2]$ and $Y_1, \ldots, Y_n$ be independent symmetric random variables satisfying $\mathbb{P}(|Y_i| \ge t) = \exp(-t^\alpha)$. Then for every vector $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and every $t \ge 0$,

$$\mathbb{P}\Big(|\sum_{i=1}^{n} a_i Y_i| \ge t\Big) \le 2\exp\Big(-c\min\Big(\frac{t^2}{\|\boldsymbol{a}\|_2^2}, \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big)$$

Proof. It is a combination of Corollaries 2.9 and 2.10 in (98).

Lemma 25 (Moments for the Sum of Weibull Distributions (Corollary 1.2 in (104))). Let $X_1, X_2, \ldots, X_n$ be a sequence of independent symmetric random variables satisfying $\mathbb{P}(|Y_i| \ge t) = \exp(-t^\alpha)$, where $0 < \alpha < 1$. Then, for $p \ge 2$ and some constant $C(\alpha)$ which depends only on α,

$$\Big\|\sum_{i=1}^{n} a_i X_i\Big\|_p \le C(\alpha)(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty).$$

Lemma 26 (Stein's Lemma (96)). Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector with joint density function $p(\boldsymbol{x})$. Suppose the score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ exists. Consider any continuously differentiable function $G(\boldsymbol{x}) : \mathbb{R}^{d_x} \to \mathbb{R}$. Then, we have

$$\mathbb{E}\Big[G(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\Big] = -\mathbb{E}\Big[\nabla_{\boldsymbol{x}} G(\boldsymbol{x})\Big].$$

Lemma 27 (Khinchin-Kahane Inequality (Theorem 1.3.1 in (105))). Let $\{a_i\}_{i=1}^n$ a finite non-random sequence, $\{\varepsilon_i\}_{i=1}^n$ be a sequence of independent Rademacher variables and $1 < p < q < \infty$. Then

$$\Big\| \sum_{i=1}^n \varepsilon_i a_i \Big\|_q \leq \Big(\frac{q-1}{p-1} \Big)^{1/2} \Big\| \sum_{i=1}^n \varepsilon_i a_i \Big\|_p .$$

Lemma 28 . Suppose each non-zero element of $\{\boldsymbol{x}_k\}_{k=1}^K$ is drawn from standard Gaussian distribution and $\|\boldsymbol{x}_k\|_0 \leq s$ for $k \in [K]$. Then we have for any $0 < \delta \leq 1$,

$$\mathbb{P}\Big(\max_{1 \leq k_1 < k_2 \leq K} |\langle \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2} \rangle| \leq C\sqrt{s}\sqrt{\log K + \log 1/\delta} \Big) \geq 1 - \delta,$$

where C is some constant.

Proof. Let us denote $\mathcal{S}_{k_1 k_2} \subset [1, 2, \ldots, p]$ as an index set such that for any $i, j \in \mathcal{S}_{k_1 k_2}$, we have $x_{k_1 i} \neq 0$ and $x_{k_2 j} \neq 0$. From the definition of $\mathcal{S}_{k_1 k_2}$, we know that $|\mathcal{S}_{k_1 k_2}| \leq s$ and $\boldsymbol{x}_{k_1}^\top \boldsymbol{x}_{k_2} = \sum_{j=1}^p x_{k_1 j} x_{k_2 j} = \sum_{j \in \mathcal{S}_{k_1 k_2}} x_{k_1 j} x_{k_2 j}$. We apply standard Hoeffding's concentration inequality,

$$\mathbb{P}\Big(|\langle \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2} \rangle| \geq t \Big) = \mathbb{P}\Big(|\sum_{j \in \mathcal{S}_{k_1 k_2}} x_{k_1 j} x_{k_2 j}| \geq t \Big) \leq e \exp\Big(- \frac{ct^2}{s} \Big).$$

Letting $ct^2/s = \log(1/\delta)$, we reach the conclusion.

Bibliography

[1] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ecm algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[2] S. Lauritzen, *Graphical Models*. Oxford Science Publications, 1996.

[3] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, pp. 19–35, 2007.

[4] J. Friedman, H. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, pp. 432–441, 2008.

[5] A. Shojaie and G. Michailidis, "Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs," *Biometrika*, vol. 97, no. 3, pp. 519–538, 2010.

[6] TCGA, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[7] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. OKelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, and TCGA, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, pp. 98–110, 2010.

[8] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[9] W. Lee and Y. Liu, "Joint estimation of multiple precision matrices with common structures," *Journal of Machine Learning Research*, vol. 16, pp. 1035–1062, 2015.

[10] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[11] H. Qiu, F. Han, H. Liu, and B. Caffo, "Joint estimation of multiple graphical models from high dimensional time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 2, pp. 487–504, 2016. [Online]. Available: http://dx.doi.org/10.1111/rssb.12123

[12] J. Wang, "Joint estimation of sparse multivariate regression and conditional graphical models," *Statistica Sinica*, vol. 25, pp. 831–851, 2015.

[13] T. T. Cai, H. Li, W. Liu, and J. Xie, "Joint estimation of multiple high-dimensional precision matrices," *Statistica Sinica*, vol. 26, no. 2, 2016.

[14] C. Peterson, F. C. Stingo, and M. Vannucci, "Bayesian inference of multiple gaussian graphical models," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 159–174, 2015.

[15] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *International ACM WWW Conference*, 2009.

[16] P. Wang, W. Sun, D. Yin, J. Yang, and Y. Chang, "Robust tree-based causal inference for complex ad effectiveness analysis," in *Proceedings of 8th ACM Conference on Web Search and Data Mining*, 2015.

[17] P. Jeziorski and I. Segal, "What makes them click: Empirical analysis of consumer demand for search advertising," *American Economic Journal*, vol. 7, pp. 24–53, 2015.

[18] Y. Chen, D. Pavlov, and J. Canny, "Large-scale behavioral targeting," in *ACM SIGKDD*, 2010.

[19] T. T. Cai, W. Liu, and H. H. Zhou, "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation," *The Annals of Statistics*, vol. 44, no. 2, pp. 455–488, 2016.

[20] W. Pan and X. Shen, "Penalized model-based clustering with application to variable selection," *Journal of Machine Learning Research*, vol. 8, pp. 1145–1164, 2007.

[21] W. Sun, J. Wang, and Y. Fang, "Regularized k-means clustering of high-dimensional data and its asymptotic consistency," *Electron. J. Statist.*, vol. 6, pp. 148–167, 2012.

[22] H. Zhou, W. Pan, and X. Shen, "Penalized model-based clustering with unconstrained covariance matrices," *Electron. J. Statist.*, vol. 3, pp. 1473–1496, 2009.

[23] C. Gao, Y. Zhu, X. Shen, and W. Pan, "Estimation of multiple networks in gaussian mixture models," *Electronic Journal of Statistics*, vol. 10, pp. 1133–1154, 2016.

[24] X. Yi and C. Caramanis, "Regularized em algorithms: A unified framework and statistical guarantees," *arXiv preprint*, vol. arXiv:1511.08551, 2015.

[25] Z. Wang, Q. Gu, Y. Ning, and H. Liu, "High dimensional em algorithm: Statistical optimization and asymptotic normality," *Advances in Neural Information Processing Systems*, pp. 2512–2520, 2015.

[26] T. Saegusa and A. Shojaie, "Joint estimation of precision matrices in heterogeneous populations," *Electronic Journal of Statistics*, p. To appear, 2016.

[27] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the em algorithm: From population to sample-based analysis," *Annals of Statistics*, 2015.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2011.

[29] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 11 2012.

[30] P. M. Kroonenberg, *Applied Multiway Data Analysis*. Wiley Series in Probability and Statistics, 2008.

[31] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, pp. 455–500, 2009.

[32] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, vol. 108, pp. 540–552, 2013.

[33] X. Li, D. Xu, H. Zhou, and L. Li, "Tucker tensor regression and neuroimaging analysis," *Statistics in Biosciences*, vol. 10, no. 3, pp. 520–545, 2018.

[34] W. W. Sun and L. Li, "Store: sparse tensor response regression and neuroimaging analysis," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4908–4944, 2017.

[35] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 355–380, 2013.

[36] S. Friedland, Q. Li, and D. Schonfeld, "Compressive sensing of sparse tensors," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4438–4447, 2014.

[37] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 208–220, 2013.

[38] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 1031–1068, 2016.

[39] A. Zhang, "Cross: Efficient low-rank tensor completion," *The Annals of Statistics*, vol. 47, no. 2, pp. 936–964, 2019.

[40] A. Montanari and N. Sun, "Spectral algorithms for tensor completion," *Communications on Pure and Applied Mathematics*, vol. 71, no. 11, pp. 2381–2425, 2018.

[41] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multi-response tensor regression," *The Annals of Statistics*, vol. to appear, 2018.

[42] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *arXiv preprint arXiv:1611.10349*, 2016.

[43] L. Li and X. Zhang, "Parsimonious tensor response regression," *Journal of the American Statistical Association*, pp. 1–16, 2017.

[44] B. Romera-Paredes, M. H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. III–1444–III–1452.

[45] J. Bien, J. Taylor, R. Tibshirani *et al.*, "A lasso for hierarchical interactions," *The Annals of Statistics*, vol. 41, no. 3, pp. 1111–1141, 2013.

[46] N. Hao and H. H. Zhang, "Interaction screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1285–1301, 2014.

[47] Y. Fan, Y. Kong, D. Li, and J. Lv, "Interaction pursuit with feature screening and selection," *arXiv preprint arXiv:1605.08933*, 2016.

[48] S. Basu, K. Kumbier, J. B. Brown, and B. Yu, "Iterative random forests to discover predictive and stable high-order interactions," *Proceedings of the National Academy of Sciences*, p. 201711236, 2018.

[49] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 517–520.

[50] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–93.

[51] W. W. Sun, J. Lu, H. Liu, and G. Cheng, "Provable sparse tensor decomposition," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 3, pp. 899–916, 2017.

[52] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra and its Applications*, vol. 523, pp. 220–262, 2017.

[53] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *The Annals of statistics*, vol. 42, no. 6, p. 2164, 2014.

[54] P.-L. Loh and M. J. Wainwright, "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *Journal of Machine Learning Research*, vol. 16, pp. 559–616, 2015.

[55] T. T. Cai and A. Zhang, "Rop: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[56] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[57] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approximation*, vol. 34, no. 1, pp. 61–88, 2011.

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

[59] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[60] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, pp. 2302–2329, 2011.

[61] E. Richard and A. Montanari, "A statistical model for tensor pca," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2897–2905.

[62] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 7311–7338, Nov 2018.

[63] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research,

E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 73–81.

[64] S. Friedland and L.-H. Lim, "Nuclear norm of higher-order tensors," *Mathematics of Computation*, vol. 87, no. 311, pp. 1255–1281, 2018.

[65] N. Ho and X. Nguyen, "Identifiability and optimal rates of convergence for parameters of multiple types in finite mixtures," *arXiv preprint arXiv:1501.02497*, 2015.

[66] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, pp. 223–232, 2012.

[67] Y. Zhu, X. Shen, and W. Pan, "Structural pursuit over multiple undirected graphs," *Journal of the American Statistical Association*, vol. 109, pp. 1683–1696, 2014.

[68] A. J. Rothman and L. Forzani, "On the existence of the weighted bridge penalized gaussian likelihood precision matrix estimator," *Electronic Journal of Statistics*, vol. 8, pp. 2693–2700, 2014.

[69] T. Cai, W. Liu, and X. Luo, "A constrained 1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 594–607, 2011.

[70] M. J. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 233–253, 2014.

[71] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," *Advances in Neural Information Processing Systems*, pp. 1260–1268, 2014.

[72] P. Bickel and E. Levina, "Covariance regularization by thresholding," *Annals of Statistics*, vol. 36, pp. 2577–2604, 2008.

[73] J. MacQueen, "Some methods for clasification and analysis of multivariate observations," *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[74] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm," *PloS one*, vol. 11, no. 9, p. e0162259, 2016.

[75] J. Wang, "Consistent selection of the number of clusters via crossvalidation," *Biometrika*, 2010.

[76] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 735–746, 2009.

[77] J. Fan, Y. Feng, and Y. Wu, "Network exploration via the adaptive lasso and scad penalties," *The Annals of Applied Statistics*, vol. 3, no. 2, p. 521, 2009.

[78] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B*, vol. 70, pp. 849–911, 2008.

[79] J. Wolfe, A. Haghighi, and D. Klein, "Fully distributed em for very large datasets," *The International Conference on Machine Learning*, pp. 1184–1191, 2008.

[80] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics., 2007.

[81] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1988.

[82] R. Vershynin, *Compressed sensing*. Cambridge Univ. Press, 2012, ch. Introduction to the non-asymptotic analysis of random matrices, pp. 210–268.

[83] C. J. Hillar and L.-H. Lim, "Most tensor problems are np-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.

[84] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.

[85] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[86] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[87] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proceedings of The 27th Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, M. F. Balcan, V. Feldman, and C. Szepesvári, Eds., vol. 35. Barcelona, Spain: PMLR, 13–15 Jun 2014, pp. 779–806.

[88] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3537–3580, 2016.

[89] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow," *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[90] M. Janzamin, H. Sedghi, and A. Anandkumar, "Score function features for discriminative learning: matrix and tensor framework," *arXiv preprint arXiv:1412.2863*, 2014.

[91] A. Anandkumar, R. Ge, and M. Janzamin, "Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates," *arXiv preprint arXiv:1402.5180*, 2014.

[92] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, April 2015.

[93] H. Hung, Y.-T. Lin, P. Chen, C.-C. Wang, S.-Y. Huang, and J.-Y. Tzeng, "Detection of gene–gene interactions using multistage sparse and low-rank regression," *Biometrics*, vol. 72, no. 1, pp. 85–94, 2016.

[94] N. D. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 757–760, 2012.

[95] S. Bubeck, *Foundations and Trends in Machine Learning*, 2015, ch. Convex Optimization: Algorithms and Complexity, pp. 231–357.

[96] C. Stein, P. Diaconis, S. Holmes, G. Reinert *et al.*, "Use of exchangeable pairs in the analysis of simulations," in *Stein's Method*. Institute of Mathematical Statistics, 2004, pp. 1–25.

[97] P. Hitczenko, S. Montgomery-Smith, and K. Oleszkiewicz, "Moment inequalities for sums of certain independent symmetric random variables," *Studia Math*, vol. 123, no. 1, pp. 15–42, 1997.

[98] M. Talagrand, "The supremum of some canonical processes," *American Journal of Mathematics*, vol. 116, no. 2, pp. 283–325, 1994.

[99] B. Yu, "Assouad, fano, and le cam," *Festschrift for Lucien Le Cam*, vol. 423, p. 435, 1997.

[100] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[101] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[102] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2005, no. 89.

[103] N. H. Nguyen, P. Drineas, and T. D. Tran, "Tensor sparsification via a bound on the spectral norm of random tensors," *Information and Inference: A Journal of the IMA*, vol. 4, no. 3, pp. 195–229, 2015.

[104] R. Bogucki, "Suprema of canonical weibull processes," *Statistics & Probability Letters*, vol. 107, pp. 253–263, 2015.

[105] V. De la Pena and E. Giné, *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.