# Sparse and Low-Rank Tensor Estimation via Cubic Sketchings

Botao Hao, Anru Zhang, and Guang Cheng

***Abstract*—In this paper, 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.**

***Index Terms*—Finite-sample analysis, non-convex optimization, tensor estimation.**

## I. INTRODUCTION

THE rapid advance in modern scientific technology gives rise to a wide range of high-dimensional tensor data [1], [2]. 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 becomes the point of interest [3]–[5]. Another example is the tensor-valued image acquisition algorithm that can considerably reduce the number of required samples by exploiting the compressibility property of signals [6], [7].

The following tensor estimation model is widely considered in recent literatures,

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

Here, $\mathscr{X}_i$ and $\epsilon_i$ 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 [8]–[15], tensor regression [3]–[5], [16]–[19], multi-task learning [20], etc.

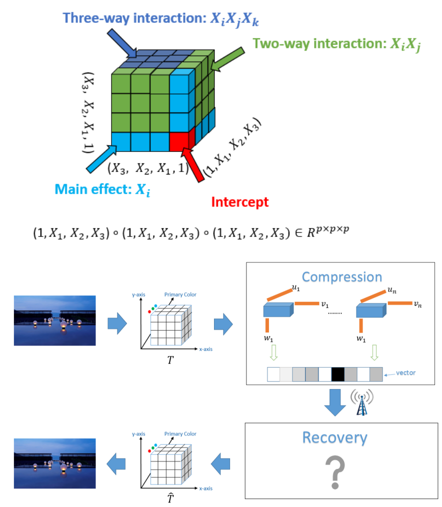


Fig. 1. Illustration for interaction reformulation and tensor image/video compression.

In this paper, we focus on the case that the measurement tensor can be written in a cubic sketching 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 whether $\mathscr{T}^*$ 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 of settings [21]–[24]. 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) 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 V.
- *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), such as hyper-spectral imaging analysis [25] and facial imaging recognition [26]. One could use Gaussian ensembles for compression such that each entry of $\mathscr{X}_i$ is i.i.d. randomly generated [3], [16], [17]. In contrast, the non-symmetric cubic sketchings, i.e., $\mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i$, reduce the memory

Manuscript received June 21, 2019; revised January 10, 2020; accepted March 11, 2020. Date of publication March 23, 2020; date of current version August 18, 2020. The work of Anru Zhang was supported in part by the NSF under Grant CAREER-1944904 and Grant DMS-1811868 and in part by the NIH under Grant R01 GM131399. The work of Guang Cheng was supported in part by the NSF under Grant DMS-1712907, Grant DMS-1811812, and Grant DMS-1821183, and in part by the Office of Naval Research (ONR) under Grant N00014-18-2759. *(Corresponding author: Anru Zhang.)*

Botao Hao is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: haobotao000@gmail.com).

Anru Zhang is with the Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: anruzhang@stat.wisc.edu).

Guang Cheng is with the Department of Statistics, Purdue University, West Lafayette, IN 47906 USA (e-mail: chengg@purdue.edu).

Communicated by M. Davenport, Associate Editor for Signal Processing.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2020.2982499

storage from $O(np_1p_2p_3)$ to $O(n(p_1+p_2+p_3))$ ($n$ is the sample size and $(p_1, p_2, p_3)$ is the tensor dimension), but still preserve the optimal statistical rate. More detailed discussions will be provided in Section VI.

In practice, the total number of measurements $n$ is considerably smaller than the number of parameters in the 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 [2] and sparsity [27]. This could highly reduce the effective dimension of the parameter and make the accurate estimation possible. Please refer to (III.2) and (VI.2) for low-rankness and sparsity assumptions.

In this paper, 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 warm start;

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

Theoretically, 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 logarithmic number of iterations, whenever $n \gtrsim K^2(s\log(ep/s))^{\frac{3}{2}}$, the proposed estimator $\widehat{\mathscr{T}}$ achieves

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

with high probability, where $s$, $K$, $p$, and $\sigma^2$ are the sparsity, rank, dimension, and noise level, respectively. We further establish the matching minimax lower bound to show that (I.2) is indeed optimal over a large class of sparse low-rank tensors. Our optimality result can be further extended to the non-sparse case (such as tensor regression [3], [17], [28], [29]) – to the best of our knowledge, this is the first statistical rate 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 [17], [30], [31], 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 [32], [33] 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 IV-C. 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 $\psi_\alpha$-norm [34], and may be of independent interest.

The literature on low-rank matrix estimation methods, e.g., the spectral method and nuclear norm minimization [35]–[37], is also related to this work. 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 [38], [39]. It is also found that simple unfolding or matricizing of tensors may lead to suboptimal results due to the loss of structural information [40]. Technically, the tensor nuclear norm is NP-hard to even approximate [9], [10], [41], and thus the method to handle tensor low-rankness is distinct from the matrix.

The rest of the paper is organized as follows. Section II provides preliminaries on notation and basic knowledge of tensor. A two-stage method for symmetric tensor estimation is proposed in Section III, with the corresponding theoretical analysis given in Section IV. A concrete application to high-order interaction effect models is described in Section V. The non-symmetric tensor estimation model is introduced and discussed in Section VI. Numerical analysis is provided in Section VII to support the proposed procedure and theoretical results of this paper. Section VIII discusses extensions to higher-order tensors. The proofs of technical results are given in supplementary materials.

## II. Preliminary

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 $l_\infty$ norm by $\max_{1\leq j\leq p}|x_j|$. Denote the set $\{1, 2, \ldots, n\}$ by $[n]$. Let $\boldsymbol{e}_j$ be the canonical vectors, whose $j$-th entry equals to 1 and all other entries equal to zero. For any two sequences $\{a_n\}_{n=1}^\infty, \{b_n\}_{n=1}^\infty$, we say $a_n = \mathcal{O}(b_n)$ if there exists some positive constant $C_0$ and sufficiently large $n_0$ such that $|a_n| \leq C_0 b_n$ for all $n \geq n_0$. We also write $a_n \asymp b_n$ if there exists $C, c > 0$ such that $ca_n \leq b_n \leq Ca_n$ for all $n \geq 1$. Additionally, $C_1, C_2, \ldots, c_1, c_2, \ldots$ are generic constants, whose actual values may be different from line to line.

We next 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 $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L] \in \mathbb{R}^{K\times L}$, their *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

$\text{vec}(\boldsymbol{X}) = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top) \in \mathbb{R}^{1\times mn}$ and column-wise $\ell_2$ norms as $\text{Norm}(\boldsymbol{X}) = (\|\boldsymbol{x}_1\|_2, \ldots, \|\boldsymbol{x}_n\|_2) \in \mathbb{R}^{1\times n}$.

In the end, we focus on tensor notation and relevant operations. Interested readers are referred to [2] 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. We say $\mathcal{X}$ is symmetric if $[\mathcal{X}]_{ijk} = [\mathcal{X}]_{ikj} = [\mathcal{X}]_{jik} = [\mathcal{X}]_{jki} = [\mathcal{X}]_{kij} = [\mathcal{X}]_{kji}$ for all $i,j,k$. Then, $\mathcal{X}$ is rank-one and symmetric if and only if it can be decomposed as $\mathcal{X} = \boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x}$ for some vector $\boldsymbol{x}$.

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

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

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 [2] with CP-rank being defined as the minimum number $K$ such that (II.1) holds. Then, $\{\boldsymbol{x}_{1k}\}_{k=1}^K, \{\boldsymbol{x}_{2k}\}_{k=1}^K, \{\boldsymbol{x}_{3k}\}_{k=1}^K$ 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

$$\|\mathcal{X}\|_F = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\sum_{k=1}^{p_3} \mathcal{X}_{ijk}^2}$$

$$\|\mathcal{X}\|_{op} := \sup_{\boldsymbol{u}\in\mathbb{R}^{p_1}, \boldsymbol{v}\in\mathbb{R}^{p_2}, \boldsymbol{w}\in\mathbb{R}^{p_3}} \frac{|\langle \mathcal{X}, \boldsymbol{u}\circ\boldsymbol{v}\circ\boldsymbol{w}\rangle|}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2\|\boldsymbol{w}\|_2}, \tag{II.2}$$

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,

$$\|\mathcal{X}\|_s := \sup_{\substack{\|\boldsymbol{a}\|=\|\boldsymbol{b}\|=\|\boldsymbol{c}\|=1 \\ \max\{\|\boldsymbol{a}\|_0, \|\boldsymbol{b}\|_0, \|\boldsymbol{c}\|_0\}\le s}} \big|\langle \mathcal{X}, \boldsymbol{a}\circ\boldsymbol{b}\circ\boldsymbol{c}\rangle\big|. \tag{II.3}$$

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 $\langle \mathcal{X}, \mathcal{Y}\rangle = (\boldsymbol{x}_1^\top \boldsymbol{y}_1)(\boldsymbol{x}_2^\top \boldsymbol{y}_2)(\boldsymbol{x}_3^\top \boldsymbol{y}_3)$.

## III. Symmetric Tensor Estimation via Cubic Sketchings

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

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

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

$$\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*,$$
$$\text{with } \|\boldsymbol{\beta}_k^*\|_2 = 1, \|\boldsymbol{\beta}_k^*\|_0 \le s, \forall k \in [K]. \tag{III.2}$$

The CP low-rankness has been widely assumed in literature for its nice scalability and simple formulation [5], [18], [25]. 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 [42], 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 [43], compressed sensing [44], matrix decomposition [45], and dictionary learning [46].

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,

$$\widehat{\mathscr{T}} = \underset{\mathscr{T}}{\text{argmin}}\, \mathcal{L}(\mathscr{T}) \quad \text{subject to } \mathscr{T} \text{ is sparse and low-rank}, \tag{III.3}$$

where

$$\mathcal{L}(\mathscr{T}) = \mathcal{L}(\eta_k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \mathscr{T}, \mathscr{X}_i\rangle)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K}\eta_k\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k\right)^3\right)^2. \tag{III.4}$$

Equivalently, (III.3) can be written as,

$$\min_{\eta_k, \boldsymbol{\beta}_k} \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3\Big)^2,$$
$$\text{s.t. } \|\boldsymbol{\beta}_k\|_2 = 1, \|\boldsymbol{\beta}_k\|_0 \le s, \text{ for } k \in [K]. \tag{III.5}$$

Clearly, (III.5) is a non-convex optimization problem. To solve it, we propose a two-stage method as described in the next two subsections.

### A. Initialization

Due to the non-convexity of (III.5), a straightforward implementation of many local search algorithms, such as gradient descent and alternating minimization, may easily get trapped into local optimums and result in sub-optimal statistical performance. Inspired by recent advances of spectral method (e.g., EM algorithm [47], phase retrieval [48], and tensor SVD [39]), 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.

**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 $\epsilon = 10^{-4}$.

**Step 1:** Calculate the moment-based tensor $\mathcal{T}_s$ as (III.6).

**Step 2:**

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

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

**Repeat** power update:

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

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

**End for**

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

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

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

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

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

Based on Lemma 4, $\mathcal{T}_s$ is an unbiased estimator of $\mathscr{T}^*$. The construction of (III.6) is motivated by the high-order Stein's identity ( [49]; also see Theorem 7 for a complete statement). Intuitively speaking, based on the third-order score function of 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 details.

**Step 2: Sparse Tensor Decomposition.** Based on the method of moment estimator obtained in Step 1, we further obtain good initialization for the factors $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ via truncation and alternating rank-1 power iterations [27], [50],

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

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, implement a clustering step, and choose the centroids as the estimates in the initialization stage. This scheme originally appears in tensor decomposition literature [43], [50], 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 first 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 updates for $l = 0, \ldots$

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

where $\times_2, \times_3$ are tensor multiplication operators defined in Section II and $T_d(\boldsymbol{x})$ is a truncation operator that sets all but the largest $d$ entries in absolute values to zero for any vector $\boldsymbol{x}$. 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, 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 $k \in [K]$.

## B. Thresholded Gradient Descent

After obtaining a warm start in the first stage, we propose to apply the thresholded gradient descent to iteratively refine the solution to the non-convex optimization problem (III.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 $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p \times K}$. Since $\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \mathcal{L}(\mathscr{T})$, we 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 A.1, $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})$ can be written as

$$\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \frac{6}{n}\Big[\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^3 \boldsymbol{\eta} - \boldsymbol{y}\Big]^\top * \Big[(\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{X}\Big]^\top, \tag{III.7}$$

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:

$$|\varphi_h(x) - x| \le h, \forall x \in \mathbb{R}, \quad \text{and } \varphi_h(x) = 0, \text{ when } |x| \le h. \tag{III.8}$$

Many widely used thresholding schemes, such as hard thresholding $H_h(x) = xI_{(|x|>h)}$, soft-thresholding $S_h(x) = \mathrm{sign}(x)\max(|x| - h, x)$, satisfy (III.8). With a slight abuse of notation, 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 Step 3, while $\boldsymbol{\eta}$ will be updated in Step 4 after finishing the update of $\boldsymbol{B}$.

**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 $\mu$, rank $K$, stopping error $\epsilon = 10^{-4}$, warm-start $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

1: **Step 3:** Let $t = 0$.
2: **Repeat** thresholded gradient descent
3: • Compute thresholding level $\boldsymbol{h}(\boldsymbol{B})$.
• Calculate the thresholded gradient descent update

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

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 (III.7).
4: **Until** $\|\boldsymbol{B}^{(T+1)} - \boldsymbol{B}^{(T)}\|_F \le \epsilon$.
5: **Step 4:** Perform column-wise normalization and update the weight as (III.10). Construct the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \widehat{\eta}_k \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k$.
6: **return** symmetric tensor estimator $\widehat{\mathscr{T}}$

**Algorithm 3** Sparse SVD

**Require:** tensor $\mathcal{T}_s$, cardinality parameter $d$.
1: Compute $\widetilde{\boldsymbol{\theta}} = T_d(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim \mathcal{N}(0, I_d)$.
2: Calculate $\boldsymbol{u}$ as the leading singular vector of $\mathcal{T}_s \times_1 \widetilde{\boldsymbol{\theta}}$.
3: **return** $T_d(\boldsymbol{u})/\|\boldsymbol{u}\|_2$

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

$$\text{vec}(\boldsymbol{B}^{(t+1)}) = \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B}^{(t)})}{\phi}}(\text{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}^{(t)}, \boldsymbol{\eta}^{(0)})). \tag{III.9}$$

Here,

- $\mu$ 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

$$\begin{aligned}\widehat{\boldsymbol{B}} &= (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_K)^\top \\ &= \Big(\frac{\boldsymbol{\beta}_1^{(T)}}{\|\boldsymbol{\beta}_1^{(T)}\|_2}, \dots, \frac{\boldsymbol{\beta}_K^{(T)}}{\|\boldsymbol{\beta}_K^{(T)}\|_2}\Big), \\ \widehat{\boldsymbol{\eta}} &= (\widehat{\eta}_1, \dots, \widehat{\eta}_K)^\top \\ &= \Big(\eta_1^{(0)}\|\boldsymbol{\beta}_1^{(T)}\|_2^3, \dots, \eta_K^{(0)}\|\boldsymbol{\beta}_K^{(T)}\|_2^3\Big)^\top. \end{aligned} \tag{III.10}$$

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):* The evaluation of the gradient (III.7) requires $\mathcal{O}(npK^2)$ operations at each iteration and can be computationally intense for large $n$ or $p$. 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 (III.7) at each iteration. This will accelerate the procedure especially in the case of large-scale settings. See Section P.2 for details.

## IV. 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.

### A. Assumptions

We first introduce the assumptions for theoretical analysis. Conditions 1-3 are on the true tensor parameter $\mathscr{T}^*$ and Conditions 4-5 are on the measurement scheme. Specifically, the first condition ensures the model identifiability for CP-decomposition.

*Condition 1 (Uniqueness of CP-Decomposition):* The CP-decomposition in (III.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 $\{1, \dots, K\}$.

For technical purposes, we introduce the following conditions to regularize the CP-decomposition of $\mathscr{T}^*$. Similar assumptions were imposed in recent tensor literature, e.g., [3], [27] and Assumption 1.1 (A4) [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

$$\|\mathscr{T}^*\|_{op} \le C\eta_{\max}^*, K = \mathcal{O}(s), R = \eta_{\max}^*/\eta_{\min}^* \le C' \tag{IV.1}$$

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 of $\boldsymbol{\beta}_k^*$.

*Remark 2:* In Condition 2, $R$ plays a similar role as a "condition number." This assumption means that the tensor is "well-conditioned," i.e., each rank-1 component is roughly of the same size.

As shown in the seminal work of [42], the estimation of low-rank tensors can be NP-hard in general. Hence, we impose the following incoherence condition.

*Condition 3 (Parameter Incoherence):* The true tensor components are incoherent such that

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

where $R$ is the singular value ratio defined in (IV.1) and $C^{''}$ is some small constant.

*Remark 3:* The preceding incoherence condition has been widely used in different scenarios in recent high-dimensional research, such as tensor decomposition [27], [50], compressed sensing [44], matrix decomposition [45], and dictionary learning [46]. It can be also viewed as a relaxation of orthogonality: if $\{\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_K^*\}$ are mutually orthogonal, $\Gamma$ equals zero. We can show from both theory (Lemma 28 in the supplementary materials) and simulation (Section VII) that the low-rank tensor $\mathscr{T}^*$ induced by (III.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 distribution.

*Condition 4 (Sub-Exponential Noise):* The noise $\{\epsilon_i\}_{i=1}^n$ are i.i.d. randomly generated with mean 0 and variance $\sigma^2$ satisfying $0 < \sigma < C\sum_{k=1}^K \eta_k^*$. $(\epsilon_i/\sigma)$ 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 is independent of $\{\mathscr{X}_i\}_{i=1}^n$.

The sample complexity condition is crucial for our algorithm especially in the initialization stage. Ignoring any poly-log factors, Condition 5 is even weaker than the sparse matrix estimation case $(n \gtrsim s^2)$ in [48].

*Condition 5 (Sample Complexity):*

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

### *B. Main Theoretical Results*

Our main Theorem 1 shows that based on a proper initializer, the output of the proposed procedure can achieve optimal estimation error rate after a sufficient number of iterations. Here, we define the 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 1 (Statistical and Optimization Errors):* Suppose Conditions 3-5 hold, $|\text{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$, and the initial estimator $\{\boldsymbol{\beta}_k^{(0)}, \eta_k^{(0)}\}_{k=1}^K$ satisfy

$$\max_{1\leq k\leq K}\left\{\left\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\right\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \lesssim K^{-1} \tag{IV.2}$$

with probability at least $1 - \mathcal{O}(1/n)$. Assume the step size $\mu \leq \mu_0$, where $\mu_0$ is defined in (A.14). Then, the output of the thresholded gradient descent update in (III.9) satisfies:

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

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

  with probability at least $1 - \mathcal{O}(tKs/n)$.
- When the total number of iterations is no smaller than

$$T^* = \left(\log(\frac{n}{\sigma^2 s\log p} \vee 1) + \log\frac{\mathcal{E}_1}{\mathcal{E}_2}\right)/\log\kappa^{-1}, \tag{IV.4}$$

  there exists a constant $C_1$ (independent of $K, s, p, n, \sigma^2$) such that the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \eta_k^{(0)}\boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)}$ satisfies

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \leq \frac{C_1\sigma^2 Ks\log p}{n} \tag{IV.5}$$

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

*Remark 4:* The error bound (IV.3) 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 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 1.

*Theorem 2 (Initialization Error):* Recall $\Gamma = \max_{1\leq k_1\neq k_2\leq K} |\langle\boldsymbol{\beta}_{k_1}^*, \boldsymbol{\beta}_{k_2}^*\rangle|$. Suppose the number of initializations $L \geq K^{C_3\gamma^{-4}}$, where $\gamma$ is a constant defined in (A.11). Given that Conditions 1-4 hold, the initial estimator obtained from Steps 1-2 with a truncation level $s \leq d \leq Cs$ satisfies

$$\max_{1\leq k\leq K}\left\{\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \leq C_2KR\delta_{n,p,s} + \sqrt{K}\Gamma^2 \tag{IV.6}$$

and

$$|\text{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$$

with probability at least $1 - 5/n$, where

$$\delta_{n,p,s} = (\log n)^3\left(\sqrt{\frac{s^3\log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}}\right). \tag{IV.7}$$

Moreover, if the sample complexity condition 5 holds, then the above bound satisfies (IV.2).

*Remark 5 (Interpretation of Initialization Error):* The upper bound of (IV.6) consists of two terms that correspond to the approximation error of $\mathcal{T}_s$ to $\mathscr{T}^*$ and the incoherence among $\boldsymbol{\beta}_k^*$'s, respectively. Especially, the former converges to zero as $n$ grows while the latter does not.

The proof of Theorems 1 and 2 are postponed to Section C-D in the supplementary materials. The combination of Theorems 1 and 2 immediately yields the following upper bound for the final estimator, which is one main result of this paper.

*Theorem 3 (Upper Bound):* Suppose Conditions 1 – 5 hold, $s \leq d \leq Cs$. After $T^*$ iterations, there exists a constant $C_1$ not depending on $K, s, p, n, \sigma^2$, such that the proposed procedure yields

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \leq \frac{C_1\sigma^2 Ks\log p}{n} \tag{IV.8}$$

with probability at least $1 - \mathcal{O}(T^*Ks/n)$, where $T^*$ is defined in (IV.4).

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

*Theorem 4 (Lower Bound):* Consider the following class of sparse and low-rank tensors,

$$\mathcal{F}_{p,K,s} = \left\{\mathscr{T}: \begin{array}{c}\mathscr{T} = \sum_{k=1}^K \eta_k\boldsymbol{\beta}_k\circ\boldsymbol{\beta}_k\circ\boldsymbol{\beta}_k, \\ \text{where } \|\boldsymbol{\beta}_k\|_0 \leq s, \text{for } k\in[K], \\ \mathscr{T}\text{satisfies Conditions 1-3}\end{array}\right\}. \tag{IV.9}$$

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 (III.1), $p \geq 20s$, and $s \geq 4$. We have the following lower bound result,

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

The proof of Theorem 4 is deferred to Section E in the supplementary materials. Combining Theorems 3 and 4, we immediately obtain the following minimax-optimal rate for sparse and low-rank tensor estimation with cubic sketchings when $\log p \asymp \log(p/s)$:

$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T}^*\in\mathcal{F}_{p,K,s}} \mathbb{E}\left\|\widetilde{\mathscr{T}}-\mathscr{T}^*\right\|_F^2 \asymp \sigma^2\frac{Ks\log(ep/s)}{n}. \quad \text{(IV.10)}$$

The rate in (IV.10) sheds light upon the effect of dimension $p$, noise level $\sigma^2$, sparsity $s$, sample size $n$ and rank $K$ to the estimation performance.

*Remark 6:* Recently, Li *et al.* [29] studied the optimal sketching for the low-rank tensor regression and gave an near-optimal sketching complexity with a sharp $(1+\varepsilon)$-worse-case error bound. Different from the framework of [29] that focuses on a deterministic setting, we study a probabilistic model with random observation noises, propose a new algorithm, and studied the minimax optimal rate of estimation errors. In addition, [5], [16], [17] considered different types of convex/non-convex algorithms for low-rank tensor regression with statistical assumptions. To our best knowledge, we are the first to achieve an optimal rate in estimation error based on polynomial-time algorithms for the tensor regression problem.

*Remark 7* (*Non-Sparse Low-Rank Tensor Estimation via Cubic-Sketchings):* When the low-rank tensor $\mathscr{T}^*$ is not necessarily sparse, i.e.,

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

we can apply the proposed procedure with all the truncation/thresholding steps removed. If $n\geq\mathcal{O}(p^{3/2})$, we can use similar arguments of Theorems 1-3 to show that the estimator $\widehat{\mathscr{T}}'$ satisfies

$$\left\|\widehat{\mathscr{T}}'-\mathscr{T}^*\right\|_F^2\lesssim\frac{\sigma^2Kp}{n} \quad \text{(IV.11)}$$

for any $\mathscr{T}^*\in\mathcal{F}_{p,K}$ with high probability. Furthermore, similar arguments of Theorem 4 imply that the rate in (IV.11) is minimax optimal.

*Remark 8* (*Comparison With Existing Matrix Results):* Our cubic sketching tensor results are far more than extensions of the existing matrix ones. For example, [32], [33] studied the low-rank matrix recovery via rank-1 projections: $y_i=\boldsymbol{x}_i^\top\boldsymbol{T}\boldsymbol{x}_i+\epsilon_i$ and proposed the convex nuclear norm minimization methods. The theoretical properties of their estimate are analyzed under a $\ell_1/\ell_2$-RIP or Restricted Uniform Boundedness condition (RUB). However, the tensor nuclear norm is computationally infeasible and one can check that our cubic sketching framework does not satisfy RIP or RUB conditions in general following the arguments in [48], [52]. 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, it requires 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 IV-C 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 analysis comparing to the ones in the matrix case [52].

### C. 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 Hoeffding's or Bernstein's concentration inequalities. Therefore, we need to develop the following 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 Lemmas 1 and 2 are given in Section B.

*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, suppose $\boldsymbol{x}_{ij}$, the $j$-th row of $\boldsymbol{X}_i$, is an isotropic sub-Gaussian vector, i.e., $\mathbb{E}\boldsymbol{x}_{ij}=0$ and $\mathrm{Cov}(\boldsymbol{x}_{ij})=I$. 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

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

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

Note that in Lemma 1, each $\boldsymbol{X}_i$ does not necessarily have independent entries, even though $\{\boldsymbol{X}_i\}_{i=1}^n$ 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 of $\mathcal{T}_s$ 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
$$\begin{aligned}\Big\|M_{\text{nsy}}-\mathbb{E}(M_{\text{nsy}})\Big\|_s&\leq C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(ep/s)}{n^2}}\\&\quad+\sqrt{\frac{s\log(ep/s)}{n}}\Big)\|\boldsymbol{\beta}_1\|_2\|\boldsymbol{\beta}_2\|_2\|\boldsymbol{\beta}_3\|_2,\end{aligned}$$
with probability at least $1-10/n^3-1/p$.
- 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
$$\begin{aligned}&\Big\|M_{\text{sym}}-\mathbb{E}(M_{\text{sym}})\Big\|_s\\&\leq C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(ep/s)}{n^2}}+\sqrt{\frac{s\log(ep/s)}{n}}\Big)\|\boldsymbol{\beta}_1\|_2^3,\end{aligned}$$
with probability at least $1-10/n^3-1/p$.

Here, $C$ is an absolute constant and $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (II.3).

## V. Application to High-Order Interaction Effect Models

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

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

for $l = 1, \ldots, n$. Here $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\eta}$ are coefficients for the main effect, pairwise interaction, and triple-wise interaction, respectively. More importantly, (V.1) can be reformulated into the following tensor form (also see the left panel of Figure 1)

$$y_l = \langle \mathcal{B}, \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \rangle + \epsilon_l, \quad l = 1, \ldots, n, \tag{V.2}$$

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:

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

We provide the following justification for assuming the tensorized coefficient $\mathcal{B}$ is low-rank and sparse. First, in modern applications, such as the biomedical research [53], the response is often driven by a small portion of coefficients and a small number of factors, leading to a highly entry-wise sparse and low-rank $\mathcal{B}$. Second, [54] suggested that it is suitable to model entry-wise sparse and low-enough rank tensors as arising from sparse loadings. Therefore, we assume $\mathcal{B}$ is CP rank-$K$ with $s$-sparse factors:

$$\mathcal{B} = \sum_{k=1}^K \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \quad \|\boldsymbol{\beta}_k\|_0 \le s,$$

where $K, s \ll p$. Then the number of parameters in (V.4), $K(p+1)$, is significantly smaller than $(p+1)^3$, the total number of parameters in the original three-way interaction effect model (V.1), which makes the consistent estimation of $\mathcal{B}$ possible in the high-dimensional case. In this case, (V.2) can be written as

$$y_l = \Big\langle \sum_{k=1}^K \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \Big\rangle + \epsilon_l, \tag{V.4}$$

where $l \in [n]$, $\|\boldsymbol{\beta}_k\|_2 = 1, \|\boldsymbol{\beta}_k\|_0 \le s, k \in [K]$.

By assuming $\boldsymbol{z}_l \overset{iid}{\sim} N_p(0, \boldsymbol{I}_p)$, the high-order interaction effect model (V.2) reduces to the symmetric tensor estimation model (III.1), except one slight difference that the first coordinate of $\boldsymbol{x}_l$, i.e., the intercept, is always 1. To accommodate this difference, we only need to adjust the initial unbiased estimate in the above two-step procedure. Let

$$\begin{aligned} \mathcal{T}_s = &\frac{1}{6n} \sum_{l=1}^n y_l \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \\ &- \frac{1}{6} \sum_{j=1}^p (\boldsymbol{a} \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{a} \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{a}), \end{aligned} \tag{V.5}$$

where $\boldsymbol{a} = \frac{1}{n}\sum_{l=1}^n y_l \boldsymbol{x}_l$. Then we construct the empirical moment-based initial tensor $\mathcal{T}_{s'}$ as

- For $i,j,k \neq 0$, $\mathcal{T}_{s'[i,j,k]} = \mathcal{T}_{s[i,j,k]}$, $\mathcal{T}_{s'[i,j,0]} = \mathcal{T}_{s[i,j,0]}$, $\mathcal{T}_{s'[0,j,k]} = \mathcal{T}_{s[0,j,k]}$, and $\mathcal{T}_{s'[i,0,k]} = \mathcal{T}_{s[i,0,k]}$.
- 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 shows that $\mathcal{T}_{s'}$ is an unbiased estimator for $\mathcal{B}$.

The theoretical results in Section IV imply the following upper and lower bounds for the three-way interaction effect estimation.

*Corollary 1:* Suppose $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ 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

$$\left\|\widehat{\mathcal{B}} - \mathcal{B}\right\|_F^2 \le C \frac{\sigma^2 K s \log p}{n} \tag{V.6}$$

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

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

Then the following lower bound holds,

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

## VI. 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

$$y_i = \langle \mathscr{T}^*, \mathscr{X}_i \rangle + \epsilon_i, \ \mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \ i \in [n], \tag{VI.1}$$

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

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

Denote

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

Then, the empirical risk function can be written compactly as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{\eta}) \\ &= \frac{1}{n} \left\| (\boldsymbol{U}^\top \boldsymbol{B}_1) * (\boldsymbol{V}^\top \boldsymbol{B}_2) * (\boldsymbol{W}^\top \boldsymbol{B}_3) \cdot \boldsymbol{\eta} - \boldsymbol{y} \right\|_2^2. \end{aligned} \tag{VI.3}$$

Since (VI.3) is non-convex but fortunately tri-convex in terms of $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, and $\boldsymbol{B}_3$, we develop a block-wise thresholded gradient descent algorithm as detailed below. The complete algorithm is deferred to Section O.1 in the supplementary materials.

**Step 1: (Method of Tensor Moments)** Construct the empirical moment-based estimator

$$\mathcal{T} := \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \in \mathbb{R}^{p_1\times p_2\times p_3} \quad \text{(VI.4)}$$

to which sparse tensor decomposition is applied for initialization.

**Step 2: (Block-wise Gradient Descent)** Lemma 17 shows that the gradient function for (VI.3) with respect to $\boldsymbol{B}_1$ can be written as

$$\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{B}_3,\boldsymbol{\eta}) = \boldsymbol{D}^\top(\boldsymbol{C}_1^\top \odot \boldsymbol{U})^\top \in \mathbb{R}^{1\times(p_1K)}, \quad \text{(VI.5)}$$

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 $\boldsymbol{B}_2^{(t)}, \boldsymbol{B}_3^{(t)}$ and update $\boldsymbol{B}_1^{(t+1)}$ via block-wise thresholded gradient descent,

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

where $\phi = \sum_{i=1}^n y_i^2/n$, $\mu$ 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 theoretical analysis for the non-symmetric case is different from the symmetric one in two folds. First, the non-symmetric cubic sketching tensor is formed by three Gaussian vectors rather than one, which leads to many differences in the calculation of high-order moments. Second, the CP-decomposition of non-symmetric tensor $\mathscr{T}^*$ (VI.2) forms a tri-convex optimization. At this point, the standard convex analysis for vanilla gradient descent [55] could be applied given a proper initialization.

With the regularity conditions detailed in Section O.1, we present the theoretical results for non-symmetric tensor estimation as follows.

*Theorem 5* (*Upper Bound):* Suppose Conditions 6 – 9 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 output of Algorithm O.1 satisfies

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

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

$$\Big\|\widehat{\mathscr{T}} - \mathscr{T}^*\Big\|_F^2 \le \mathcal{O}_p\Big(\frac{\sigma^2 Ks\log p_0}{n}\Big).$$

*Theorem 6* (*Lower Bound):* Consider the class of incoherent sparse and low-rank tensors $\mathcal{F} = \{\mathscr{T} : \mathscr{T} = \sum_{k=1}^K \boldsymbol{\beta}_{1k}\circ\boldsymbol{\beta}_{2k}\circ\boldsymbol{\beta}_{3k}, \|\boldsymbol{\beta}_{i,k}\|_0 \le s \text{ for } i = 1,2,3, k = 1,\ldots,K\}$. If $\{\mathscr{X}_i\}_{i=1}^n$ are i.i.d standard normal cubic sketchings, $\epsilon \overset{iid}{\sim} N(0,\sigma^2)$, $\min\{p_1,p_2,p_3\} \ge 20s$, and $s\ge 4$, we have

$$\inf_{\widehat{\mathscr{T}}}\sup_{\mathscr{T}\in\mathcal{F}}\mathbb{E}\Big\|\widehat{\mathscr{T}} - \mathscr{T}\Big\|_F^2 \ge \frac{C\sigma^2 sK\log(e\cdot p_0/s)}{n}. \quad \text{(VI.6)}$$

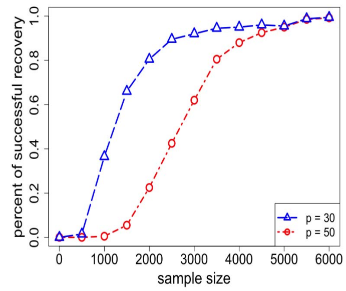


Fig. 2. Successful rate of recovery with varying sample size.

Theorems 5 and 6 imply that the proposed algorithm achieves a minimax-optimal rate of estimation error in the class of $\mathcal{F}$ as long as $\log(p_0) \asymp \log(p_0/s)$.

## VII. Numerical Results

In this section, we investigate the effect of noise level, CP-rank, sample size, dimension, and sparsity on the estimation performance by simulation studies. We also investigate the numerical performance of the proposed algorithm when the incoherence assumption required in the theoretical analysis fails to hold.

In each setting, we generate $\mathscr{T}^* = \sum_{k=1}^K \boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*$, where $|\mathrm{supp}(\boldsymbol{\beta}_k^*)| = s$, the support of $\boldsymbol{\beta}_k^*$ is uniformly selected from $\{1,\ldots,p\}$, and the nonzero entries of $\boldsymbol{\beta}_k^*$ are drawn randomly from standard normal distribution. Then, we calculate $\eta_k^* \leftarrow \|\boldsymbol{\beta}_k^*\|_2^3$ and normalize $\boldsymbol{\beta}_k^* \leftarrow \boldsymbol{\beta}_k^*/\|\boldsymbol{\beta}_k^*\|_2$. The cubic sketchings $\{\mathscr{X}_i\}_{i=1}^n$ are generated as $\mathscr{X}_i = \boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i$ and $\boldsymbol{x}_i \overset{iid}{\sim} N(0,1)$. The noise satisfies $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0,\sigma^2)$ or Laplace$(0,\sigma/\sqrt{2})$. Additionally, we adopt the following stopping rules in iterations: (1) the initialization iteration (Step 2 in Algorithm 1) is stopped if $\|\boldsymbol{b}_m^{(l+1)} - \boldsymbol{b}_m^{(l)}\|_2 \le 10^{-6}$; (2) the gradient update iteration (Step 3 in Algorithm 2) is stopped if $\|\boldsymbol{B}^{(T+1)} - \boldsymbol{B}^{(T)}\|_F \le 10^{-6}$. The numerical results are based on 200 repetitions unless otherwise specified. The code was written in R and implemented on an Intel Xeon-E5 processor with 64 GB of RAM.

First, we consider the percentage 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 $2.7\times10^4$ or $1.25\times10^5$. The sample size $n$ ranges from 500 to 6000. Each recovery is called "successful" if the relative error $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F/\|\mathscr{T}^*\|_F < 10^{-4}$. We report the average successful recovery rate in Figure VII. We can see from Figure VII that the empirical relation among successful recovery, dimension, and sample size is consistent with the theoretical results in Section IV.

We then move to the noisy case. Select $K = 3$, $s/p = 0.3$, $p \in \{30, 50\}$, $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0,\sigma^2)$. We consider two scenarios: (1) sample size $n$ = 6000, 8000, or 10000, $s/p = 0.3$, the noise level $\sigma$ 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 $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F/\|\mathscr{T}^*\|_F$ in these two scenarios are plotted in Figures 3 and 4, respectively. These results show that the proposed procedure achieves a

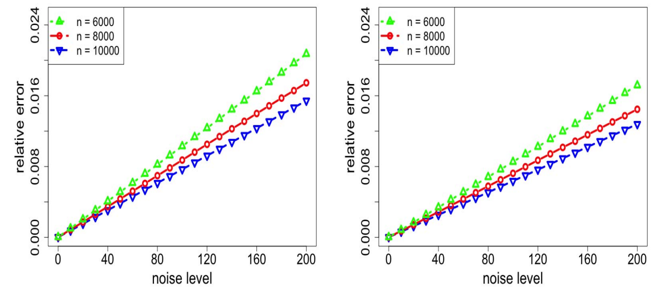


Fig. 3. Estimation error under different noise levels. Left panel: $p = 30$, right panel: $p = 50$.

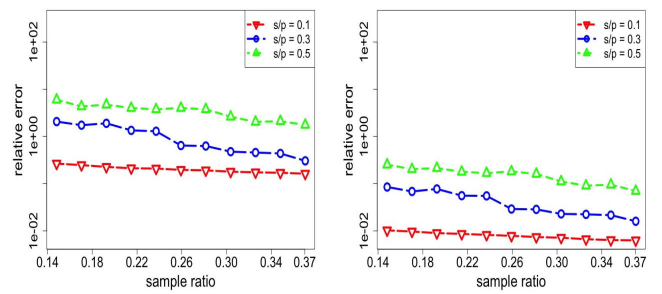


Fig. 4. Estimation error under different dimension/sample ratio ($n/p^3$). Left panel: initial estimation error, right panel: final estimation error.

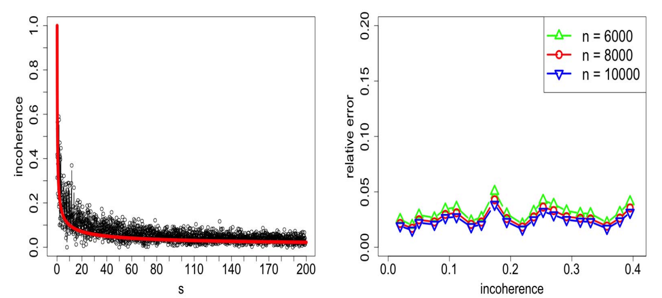


Fig. 5. Left panel: incoherence parameter $\Gamma$ with varying sparsity. Here, the red line corresponds to the rate $\sqrt{s}$ required in the theoretical analysis. Right panel: average relative estimation error for tensors with varying incoherence.

good performance – Algorithms 1 and 2 yield more accurate estimation with smaller variance $\sigma^2$ and/or large value of sample size $n$.

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 $\Gamma$ defined in Condition 3. The left panel of Figure 5 shows that the incoherence parameter $\Gamma$ decays in a polynomial rate as $s$ grows, which matches the bound in Condition 3. Recall a theoretical justification on this point is also provided in Lemma 28.

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 $\Gamma_j$ defined in Condition 3, then manually pick

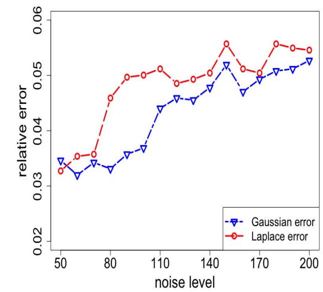


Fig. 6. Comparison of estimation errors between Laplace error and Gaussian error.

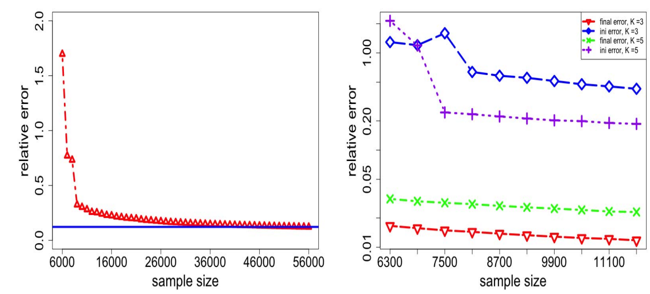


Fig. 7. Log relative estimation error of initial estimation error (left panel) and initialization/final estimation error (right panel).

40 $\mathscr{T}_{j'}^*$ such that

$$0.01 \cdot (j' - 1) \leq \Gamma_{j'} \leq 0.01 \cdot j' \quad \text{for} \quad j' = \{1, 2, \ldots, 40\}.$$

In this way, we obtain a set of tensor parameters $\{\mathscr{T}_{j'}^*\}$ with incoherence uniformly varying from 0 to 0.4. The right panel of Figure 5 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 a setting with Laplacian noise. 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 $\sigma$, the average estimation error and its comparison with Gaussian noise setting are provided in Figure 6. We note that the estimation errors under Laplace noise are slightly higher than those under Gaussian noise.

We also compare the estimation errors of initial and final estimators for different ranks and sample sizes. Set $K = 3$, $p = 30, s/p = 0.3$ and consider the noiseless setting. It is clear from Figure 7 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 2: as discussed in Remark 5, the initial stage may yield an inconsistent estimator due to the incoherence among $\boldsymbol{\beta}_k$'s. We also evaluate and compare the estimation errors for both initial and final estimators. From the right panel of Figure 7, we can see that the final estimator is more stable and accurate compared to the initial one, which illustrates the merit of thresholded gradient descent step of the proposed procedure.

Finally, we compare the performance of the proposed method with the alternating least square (ALS)-based tensor regression method [3]. We specifically consider two schemes for the initialization of ALS: (a) $\{\boldsymbol{\beta}_k^{(0)}\}$ are i.i.d. standard

TABLE I
ESTIMATION ERROR AND STANDARD DEVIATION (IN SUBSCRIPT) OF THE PROPOSED METHOD AND ALS-BASED METHOD

| Sample size | ours | warm start | cold start | initial |
|---|---|---|---|---|
| $n = 4000$ | $4.02_{0.13}$ | $32.82_{1.79}$ | $37.78_{1.23}$ | $38.03_{1.74}$ |
| $n = 5000$ | $1.94_{0.09}$ | $32.34_{2.34}$ | $36.96_{2.10}$ | $33.71_{1.78}$ |
| $n = 6000$ | $1.77_{0.09}$ | $22.22_{1.21}$ | $59.97_{3.40}$ | $25.57_{1.48}$ |

Gaussian (cold start), and (b) $\{\boldsymbol{\beta}_k^{(0)}\}$ are 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 apply both the proposed procedure and the ALS-based algorithm and record the average estimation errors with standard deviations for both initial and final estimators. From the result in Table VII, one can see the proposed algorithm significantly outperforms the ALS under both cold and warm start schemes. The main reason is pointed out in Remark 8: the cubic sketching setting possesses distinct aspects compared with the i.i.d. random Gaussian sketching setting, so that the method proposed by [3] does not exactly fit here.

## VIII. DISCUSSION

This paper focuses on the third order tensor estimation via cubic sketchings. Moreover, all results can be extended to the higher-order case via high-order sketchings. To be specific, suppose

$$y_i = \langle \mathscr{T}^*, \boldsymbol{x}_i^{\otimes m} \rangle + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\mathscr{T}^* \in (\mathbb{R}^p)^{\otimes m}$ is an order-$m$, sparse, and low-rank tensor. In order to estimate $\mathscr{T}^*$ based on $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, one can first construct the order-$m$ moment-based estimator using a generalized version of Theorem 7 and the fact that the score functions $\mathcal{S}_m(\boldsymbol{x}) = (-1)^m \nabla^m p(\boldsymbol{x})/p(\boldsymbol{x})$ for 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, we can show if mild conditions hold and $n \geq C(\log n)^m (s \log p)^{m/2}$, the proposed procedure achieves

$$\left\| \widetilde{\mathscr{T}} - \mathscr{T}^* \right\|_F^2 \lesssim \sigma^2 \frac{Kms \log(p/s)}{n}$$

with high probability. The minimax optimality can be shown similarly.

## APPENDIX

This appendix contains five parts: (1) Sections A-B provide detailed proofs for empirical moment estimator and concentration results; (2) Sections C-N provide additional proofs for the main theoretical results of this paper; (4) Section O covers the pseudo-code, conditions and main proofs of non-symmetric tensor estimation; (5) Section P discusses the matrix form of gradient function and stochastic gradient descent; (6) Section Q provides several technical lemmas and their proofs.

### A. Moment Calculation

We first introduce three lemmas to show that the empirical moment based tensors (III.6), (V.5), and (VI.4) are all unbiased estimators for the target low-rank tensor in the corresponding scenarios. Detail proofs of three lemmas are postponed to Sections G.1, G.2 and G.3 in the supplementary materials.

*Lemma 3 (Unbiasedness of Moment Estimator Under Non-Symmetric Sketchings):* For non-symmetric tensor estimation model (VI.1) & (VI.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 [49]. To fix the idea, we present only third order result for simplicity. The extension to higher-order is straightforward.

*Theorem 7 (Third-Order Stein's Identity, [49]):* 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

$$\mathbb{E}\left[G(\boldsymbol{x}) \cdot \mathcal{S}_3(\boldsymbol{x})\right] = \mathbb{E}\left[\nabla^3 G(\boldsymbol{x})\right]. \quad \text{(A.1)}$$

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

$$\mathcal{S}_m(\boldsymbol{x}) := -\mathcal{S}_{m-1}(\boldsymbol{x}) \circ \nabla \log p(\boldsymbol{x}) - \nabla \mathcal{S}_{m-1}(\boldsymbol{x}), \quad \text{(A.2)}$$

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 [56]). The proof of Theorem 7 relies on iteratively applying the recursion representation of score function (A.2) and the first-order Stein's lemma (Lemma 26). We provide the detailed proof in Section F for the sake of completeness.

In particular, if $\boldsymbol{x}$ follows a standard Gaussian vector, each order score function can be calculated based on (A.2) as follows,

$$\begin{aligned} &\mathcal{S}_1(\boldsymbol{x}) = \boldsymbol{x}, \mathcal{S}_2(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{x} - I_{d \times d}, \\ &\mathcal{S}_3(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x} \\ &- \sum_{j=1}^p \Big( \boldsymbol{x} \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{x} \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{x} \Big). \end{aligned} \quad \text{(A.3)}$$

Interestingly, if we let $G(\boldsymbol{x}) = \sum_{k=1}^{K} \eta_k^*(\boldsymbol{x}^\top \boldsymbol{\beta}_k^*)^3$, then

$$\frac{1}{6}\nabla^3 G(\boldsymbol{x}) = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*, \tag{A.4}$$

which is exactly $\mathscr{T}^*$. Connecting this fact with (A.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 (III.1) & (IV.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 (A.3). By using the conclusion in Theorem 7 and the fact (A.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 $\epsilon_i$ is independent of $\boldsymbol{x}_i$. This ends the proof. ∎

Although the interaction effect model (V.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 (V.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.,

$$\mathbb{E}(\mathcal{T}_{s'}) = \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k.$$

### *B. 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 $\psi_\alpha$-norm.

*Definition 1 ($\psi_\alpha$-Norm [34]):* The $\psi_\alpha$-norm of any random variable $X$ and $\alpha > 0$ is defined as

$$\|X\|_{\psi_\alpha} := \inf\Big\{C \in (0,\infty): \ \mathbb{E}[\exp(|X|/C)^\alpha] \le 2\Big\}.$$

Particularly, a random variable who has a bounded $\psi_2$-norm or bounded $\psi_1$-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 $\psi_\alpha$-norm.

*Lemma 6:* Suppose $X_1, \ldots, X_n$ are $n$ independent random variables satisfying $\|X_i\|_{\psi_\alpha} \le b$ with $\alpha > 0$, then for all $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $p \ge 2$,

$$\Big(\mathbb{E}\Big|\sum_{i=1}^{n} a_i X_i - \mathbb{E}(\sum_{i=1}^{n} a_i X_i)\Big|^p\Big)^{\frac{1}{p}} \le \begin{cases} C_1(\alpha)b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty\big), & \text{if } 0<\alpha<1; \\ C_2(\alpha)b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\big), & \text{if } \alpha \ge 1. \end{cases} \tag{A.5}$$

where $1/\alpha^* + 1/\alpha = 1$, $C_1(\alpha), C_2(\alpha)$ are some absolute constants only depending on $\alpha$.

If $0 < \alpha < 1$, (A.5) is a combination of Theorem 6.2 in [57] and the fact that the $p$-th moment of a Weibull variable with parameter $\alpha$ is of order $p^{1/\alpha}$. If $\alpha \ge 1$, (A.5) follows from a combination of Corollaries 2.9 and 2.10 in [58]. Continuing with standard symmetrization arguments, we reach the conclusion for general random variables. When $\alpha = 1$ or 2, (A.5) coincides with standard moment bounds for a sum of sub-Gaussian and sub-exponential random variables in [59]. The detailed proof of Lemma 6 is postponed to Section H.

When $0 < \alpha < 1$, by Chebyshev's inequality, one can obtain the following exponential tail bound for the sum of random variables with bounded $\psi_\alpha$-norm. This lemma generalizes the Hoeffding-type concentration inequality for sub-Gaussian random variables (see, e.g. Proposition 5.10 in [59]), and Bernstein-type concentration inequality for sub-exponential random variables (see, e.g. Proposition 5.16 in [59]).

*Lemma 7:* Suppose $0 < \alpha < 1$, $X_1, \ldots, X_n$ are independent random variables satisfying $\|X_i\|_{\psi_\alpha} \le b$. Then there exists absolute constant $C(\alpha)$ only depending on $\alpha$ such that for any $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $0 < \delta < 1/e^2$,

$$\Big|\sum_{i=1}^{n} a_i X_i - \mathbb{E}(\sum_{i=1}^{n} a_i X_i)\Big| \le C(\alpha)b\|\boldsymbol{a}\|_2(\log\delta^{-1})^{1/2} + C(\alpha)b\|\boldsymbol{a}\|_\infty(\log\delta^{-1})^{1/\alpha},$$

with probability at least $1 - \delta$.

*Proof:* For any $t > 0$, by Markov's inequality,

$$\begin{aligned}&\mathbb{P}\Big(\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big)\Big| \geq t\Big)\\ =&\mathbb{P}\Big(\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big)\Big|^p \geq t^p\Big)\\ \leq&\frac{\mathbb{E}\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big)\Big|^p}{t^p}\\ \leq&\frac{C(\alpha)^p b^p\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty\Big)^p}{t^p},\end{aligned}$$

where the last inequality is from Lemma 6. We set $t$ such that $\exp(-p) = C(\alpha)^p b^p(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty)^p/t^p$. Then for $p \geq 2$,

$$\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big)\Big| \leq eC(\alpha)b\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty\Big)$$

holds with probability at least $1 - \exp(-p)$. Letting $\delta = \exp(-p)$, we have that for any $0 < \delta < 1/e^2$,

$$\begin{aligned}&\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big)\Big|\\ \leq&C(\alpha)b\Big(\|\boldsymbol{a}\|_2(\log\delta^{-1})^{1/2} + \|\boldsymbol{a}\|_\infty(\log\delta^{-1})^{1/\alpha}\Big),\end{aligned}$$

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 $\psi_\alpha$-norm.

*Lemma 8 ($\psi_\alpha$ for Product of Random Variables):* Suppose $X_1, \ldots, X_m$ are $m$ random variables (not necessarily independent) with $\psi_\alpha$-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

$$\begin{aligned}&\exp\Big(|\prod_{j=1}^m \frac{x_j}{K_j}|\Big)^{\alpha/m} \leq \exp\Big(\frac{1}{m}\sum_{j=1}^m |\frac{x_j}{K_j}|^\alpha\Big)\\ &= \Big(\prod_{j=1}^m \exp(|\frac{x_j}{K_j}|^\alpha)\Big)^{1/m} \leq \frac{1}{m}\sum_{j=1}^m \exp\Big(|\frac{x_j}{K_j}|^\alpha\Big). \quad \text{(A.6)}\end{aligned}$$

From the definition of $\psi_\alpha$-norm, for $j = 1, 2, \ldots, m$, each individual $X_j$ has

$$\mathbb{E}\Big(\exp(\frac{|X_j|}{K_j})^\alpha\Big) \leq 2. \quad \text{(A.7)}$$

Putting (A.6) and (A.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 $\psi_2$-norm of $\boldsymbol{X}_j^\top\boldsymbol{\beta}_j$ is bounded by $\|\boldsymbol{\beta}_j\|_2$ [59]. 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,

$$\begin{aligned}\mathcal{G}_n &= \{\max_i\{|x_{1i}|, |x_{2i}|, |x_{3i}|\} \leq M\},\\ \mathcal{G} &= \{\max\{|x_1|, |x_2|, |x_3|\} \leq M\},\end{aligned}$$

where $M$ is a large constant to be specified later. Let $\|M_{\text{nsy}} - \mathbb{E}(M_{\text{nsy}})\|_s$ upper bounded by $M_1 + M_2$ where

$$\begin{aligned}M_1 = \Big\|&\frac{1}{n}\sum_{i=1}^n x_{1i}x_{2i}x_{3i}\boldsymbol{x}_{1i}\circ\boldsymbol{x}_{2i}\circ\boldsymbol{x}_{3i}\\ &-\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3|\mathcal{G}\Big)\Big\|_s\end{aligned}$$

and

$$M_2 = \Big\|\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3|\mathcal{G}\Big) - \boldsymbol{e}_1\circ\boldsymbol{e}_1\circ\boldsymbol{e}_1\Big\|_s.$$

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

$$\mathbb{P}\Big(M_1 \geq t\Big) \leq \mathbb{P}\Big(\mathcal{G}_n^c\Big) + \mathbb{P}\Big(M_{11} \geq t\Big),$$

where

$$M_{11} = \Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}'\boldsymbol{x}_{1i}' \circ x_{3i}'\boldsymbol{x}_{2i}' \circ x_{i1}'\boldsymbol{x}_{3i}' - \mathbb{E}\Big(x_1'\boldsymbol{x}_1' \circ x_2'\boldsymbol{x}_2' \circ x_3'\boldsymbol{x}_3'\Big)\Big\|_s.$$

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 $\psi_2$-norm $M^2$. Applying Lemma 20, we obtain

$$\mathbb{P}\Big(M_{11} \geq C_1 M^6 \delta_{n,s}\Big) \leq \frac{1}{p},$$

where $\delta_{n,s} = ((s\log(p/s))^3/n^2)^{1/2} + (s\log(p/s)/n)^{1/2}$.

On the other hand,

$$\mathbb{P}(\mathcal{G}_n^c) \leq 3\sum_{i=1}^n \mathbb{P}(|x_{1i}| \geq M) \leq 3ne^{1-C_2M^2}$$

Putting the above bounds together, we obtain

$$\mathbb{P}\Big(M_1 \geq C_1\ M^6\delta_{n,s}\Big) \leq \frac{1}{p} + 3ne^{1-C_2M^2}.$$

By setting $M = 2\sqrt{\log n/C_2}$, the bound of $M_1$ reduces to

$$\mathbb{P}\Big(M_1 \geq \frac{64C_1}{C_2^3}\delta_{n,s}(\log n)^3\Big) \leq \frac{1}{p} + \frac{3e}{n^3}. \tag{A.8}$$

**Bounding** $M_2$**.** From the definitions of $M_2$ and sparse spectral norm,

$$\begin{aligned} M_2 &= \Big\|\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1 \circ \boldsymbol{x}_2 \circ \boldsymbol{x}_3|\mathcal{G}\Big) - \boldsymbol{e}_1 \circ \boldsymbol{e}_1 \circ \boldsymbol{e}_1\Big\|_s \\ &= \sup_{\mathcal{D}(a,b,c)} \Big|\mathbb{E}\Big(x_1x_2x_3(\boldsymbol{x}_1^\top \boldsymbol{a})(\boldsymbol{x}_2^\top \boldsymbol{b})(\boldsymbol{x}_3^\top \boldsymbol{c})|\mathcal{G}\Big) - a_1b_1c_1\Big|. \end{aligned}$$

where

$$\mathcal{D} = \Big\{\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2 = \|\boldsymbol{c}\|_2 = 1, \max\{\|\boldsymbol{a}\|_0, \|\boldsymbol{b}\|_0, \|\boldsymbol{c}\|_0\} \leq s\Big\}.$$

Since $x_{1j}$ is independent of $x_{1k}$ for any $j \neq k$, $\mathbb{E}(x_1(\boldsymbol{x}_1^\top \varrho\boldsymbol{a})|\mathcal{G}) = \mathbb{E}(x_1^2 a_1|\mathcal{G})$. Similar results hold for $\boldsymbol{x}_2, \boldsymbol{x}_3$. Then we have

$$\begin{aligned} M_2 &= \sup_{\mathcal{D}} |a_1b_1c_1|\Big|\mathbb{E}\Big(x_1^2x_2^2x_3^2|\mathcal{G}\Big) - 1\Big| \\ &\leq \Big|\mathbb{E}\Big(x_1^2x_2^2x_3^2|\mathcal{G}\Big) - 1\Big| = \Big|\mathbb{E}\Big(x_1^2\Big||x_1| \leq M\Big) \\ &\quad \mathbb{E}\Big(x_2^2\Big||x_2| \leq M\Big)\mathbb{E}\Big(x_3^2\Big||x_3| \leq M\Big) - 1\Big|. \end{aligned}$$

By the basic property of Gaussian random variable, we can show

$$1 \geq \mathbb{E}\Big(x_i^2\Big||x_i| \leq M\Big) \geq 1 - 2Me^{-M^2/2}, \quad i = 1,2,3.$$

Plugging them into $M_2$, we have

$$\begin{aligned} M_2 &\leq \Big|\Big(1 - 2Me^{-M^2/2}\Big)^3 - 1\Big| \\ &\leq \Big|12M^2e^{-M^2} - 6Me^{-M^2/2} - 8M^3e^{-3M^2/2}\Big| \\ &\leq \Big|26M^3e^{-M^2/2}\Big|, \end{aligned}$$

where the last inequality holds for a large $M > 0$. By the choice of $M = 2\sqrt{\log n/C_2}$, we have $M_2 \leq 208/C_2^{3/2}(\log n)^{\frac{3}{2}}/n^2$ for some constant $C_2$. When $n$ is large, this rate is negligible comparing with (A.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

$$M_1 + M_2 \leq C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big),$$

with probability at least $1 - 10/n^3 - 1/p$. This concludes the proof of non-symmetric part. The proof of symmetric part remains similar and thus is omitted here. ∎

### C. Proof of Theorem 2: Initialization Effect

Theorem 2 gives an approximation error upper bound for the sparse-tensor-decomposition-based initial estimator. In Step I of Section III-A, the original problem can be reformatted to a version of tensor denoising:

$$\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}, \quad \text{where} \quad \mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s). \tag{A.9}$$

The key difference between our model (A.9) and recent works [27], [50] is that $\mathcal{E}$ arises from empirical moment approximation, rather than the random observation noise considered in [50] and [27]. Next lemma gives an upper bound for the approximation error. The proof of Lemma 9 is deferred to Section I.

*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 (III.6). Suppose Condition 4 is satisfied and $s \leq d \leq Cs$. Then

$$\|\mathcal{E}\|_{s+d} \leq 2C_1\sum_{k=1}^K \eta_k^*\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big)(\log n)^4, \tag{A.10}$$

with probability at least $1 - 5/n$ for some uniform constant $C_1$.

Next we denote the following quantity for simplicity,

$$\gamma = C_2\min\Big\{\frac{R^{-1}}{6} - \frac{\sqrt{K}}{s}, \frac{R^{-1}}{4\sqrt{5}} - \frac{2}{\sqrt{s}}\Big(1+\sqrt{\frac{K}{s}}\Big)^2\Big\}, \tag{A.11}$$

where $R$ is the singular value ratio, $K$ is the CP-rank, $s$ is the sparsity parameter, $\Gamma$ 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 (A.9) 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 $N \geq C_4\log\Big(\gamma/\Big(\frac{1}{\eta_{\min}^*}\|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2\Big)\Big)$ for constants $C_3, C_4$, the truncation parameter $s \leq d \leq Cs$. Then the sparse-tensor-decomposition-based initialization satisfies

$$\max\left\{\|\boldsymbol{\beta}_k^{(0)}-\boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)}-\eta_k^*|\right\} \leq \frac{C_4}{\eta_{\min}^*}\|\mathcal{E}\|_{s+d}+\sqrt{K}\Gamma^2, \tag{A.12}$$

for any $k \in [K]$.

The proof of Lemma 10 essentially follows Theorem 3.9 in [27], we thus omit the detailed proof here. The upper bound in (A.12) 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 (A.10) and Condition 2, the right side of (A.12) reduces to

$$\frac{C_4}{\eta_{\min}^*}\|\mathcal{E}\|_{s+d}+\sqrt{K}\Gamma^2 \leq 2\ C_1C_4KR\Big(\sqrt{\tfrac{s^3\log^3(p/s)}{n^2}} +\sqrt{\frac{s\log(p/s)}{n}}\Big)(\log n)^4+\sqrt{K}\Gamma^2,$$

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 $1-5/n$. ■

### D. Proof of Theorem 1: 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 J. 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

$$\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \leq 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2,\ \max_{k\in[K]}\left|\eta_k-\eta_k^*\right|\leq\varepsilon_0, \tag{A.13}$$

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 $\mu$ satisfies

$$0<\mu\leq\mu_0=\frac{32R^{-20/3}}{3K[220+270K]^2}, \tag{A.14}$$

then $\{\boldsymbol{\beta}_k^{(t+1)}\}$ can be upper bounded as

$$\begin{aligned}&\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t+1)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\ \leq&\underbrace{\Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2}_{\text{optimization error}}\\ &\underbrace{+2C_0\mu^2K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2 s\log p}{n}}_{\text{statistical error}},\end{aligned}$$

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

In order to apply Lemma 11, we prove that the required condition (A.13) holds at every iteration step $t$ by induction. When $t=0$, by (IV.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

$$\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\leq\sum_{k=1}^K 4\eta_k^{*\frac{2}{3}}\varepsilon_0^2\leq 4\ K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2,$$

with probability at least $1-\mathcal{O}(K/n)$, which means (A.13) holds for $t=0$.

Suppose (A.13) holds at the iteration step $t-1$, which implies

$$\begin{aligned}&\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\ \leq&\Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^K\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t-1)}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\ &+\mu 2C_0K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*\frac{4}{3}}\frac{\sigma^2 s\log p}{n}\\ \leq&4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2-\mu\Big(128KR^{-\frac{8}{3}}\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2\\ &-2C_0K^{-2}R^{-\frac{8}{3}}\eta_{\min}^{*\frac{4}{3}}\frac{\sigma^2 s\log p}{n}\Big).\end{aligned}$$

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, (A.13) 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 $t = 1, 2, \ldots,$

$$\begin{aligned}&\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\ &\leq \left(1 - 32\mu K^{-2}R^{-\frac{8}{3}}\right)^t \sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\ &\quad + \frac{C_0\eta_{\min}^{*-\frac{4}{3}}}{16}\frac{\sigma^2 s\log p}{n},\end{aligned}$$

with probability at least $1 - \mathcal{O}(tKs/n)$. This concludes the first part of Theorem 1.

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

$$\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(T^*)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2 \leq \frac{C_3\eta_{\min}^{*-\frac{4}{3}}}{8}\frac{\sigma^2 s\log p}{n}, \quad \text{(A.15)}$$

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 K.

*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

$$\begin{aligned}&\left\|\mathscr{T} - \mathscr{T}^*\right\|_F^2\\ &\leq 9(1+c)\Big(\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\Big)\Big(\sum_{k=1}^{K}(\sqrt[3]{\eta_k^*})^4\Big)\end{aligned}$$

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 (A.15) and Lemma 12, we have

$$\begin{aligned}\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 &\leq 9(1+\varepsilon_0)\frac{C_3\eta_{\min}^{*-\frac{4}{3}}}{8}\frac{\sigma^2 s\log p}{n}K\eta_{\max}^{*\frac{4}{3}},\\ &= \frac{9C_3R}{4}\frac{\sigma^2 Ks\log p}{n},\end{aligned}$$

with probability at least $1 - \mathcal{O}(TKs/n)$. By setting $C_1 = 9C_2/4$, we complete the proof of Theorem 1. ■

## *E. Proofs of Theorems 4 and 6: Minimax Lower Bounds*

We first consider the proof for Theorem 6 on non-symmetric tensor estimation. Without loss of generality we assume $p = \max\{p_1, p_2, p_3\}$. We uniformly randomly generate $\{\Omega^{(k,m)}\}_{\substack{m=1,\ldots,M\\k=1,\ldots,K}}$ as $MK$ subsets of $\{1,\ldots,p\}$ with cardinality of $s$. Here $M > 0$ is a large integer to be specified later. Then we construct $\{\boldsymbol{\beta}^{(k,m)}\}_{\substack{m=1,\ldots,M\\k=1,\ldots,K}} \subseteq \mathbb{R}^p$ 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, $\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \leq 2s\lambda$ for any $1\leq k\leq K$, $1\leq m_1, m_2\leq M$. 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

$$w^{(k,m_1,m_2)} = \left|\Omega^{(k,m_1)}\cap\Omega^{(k,m_2)}\right|, \qquad \text{(A.16)}$$

then for any $s/2\leq t\leq s$,

$$\begin{aligned}\mathbb{P}\left(w^{(k,m_1,m_2)} = t\right) &= \frac{\frac{s\cdots(s-t+1)}{t!}\cdot\frac{(p-s)\cdots(p-2s+t+1)}{(s-t)!}}{\frac{p\cdots(p-s+1)}{s!}}\\ &\leq \binom{s}{t}\cdot\left(\frac{s}{p-s+1}\right)^t\\ &\leq 2^s\left(\frac{s}{p-s+1}\right)^t \leq \left(\frac{4s}{p-s+1}\right)^t.\end{aligned}$$

Thus, if $\eta > 0$, the moment generating function of $w^{(k,m_1,m_2)} - \frac{s}{2}$ satisfies

$$\begin{aligned}&\mathbb{E}\exp\left(\eta\left(w^{(k,m_1,m_2)} - \frac{s}{2}\right)\right)\\ &\leq \exp(0)\cdot\mathbb{P}\left(w^{(k,m_1,m_2)} \leq \frac{s}{2}\right)\\ &\quad + \sum_{t=\lfloor s/2\rfloor+1}^{s}\exp\left(\eta\left(t - \frac{s}{2}\right)\right)\cdot\mathbb{P}\left(w^{(k,m_1,m_2)} = t\right)\\ &\leq 1 + \sum_{t=\lfloor s/2\rfloor+1}^{s}(4s/(p-s+1))^t\exp\left(\eta(t - s/2)\right)\\ &= 1 + \left(\frac{4s}{p-s+1}\right)^{\lfloor s/2\rfloor+1}\\ &\quad \sum_{t=0}^{s-\lfloor s/2\rfloor-1}\left(\frac{4s}{p-s+1}\right)^t\exp\left(\eta(t+\lfloor s/2\rfloor+1-s/2)\right)\\ &\overset{(*)}{\leq} 1 + \left(\frac{4s}{p-s+1}\right)^{s/2}\sum_{t=0}^{s-\lfloor s/2\rfloor-1}\left(\frac{4se^\eta}{p-s+1}\right)^t\\ &= 1 + \left(\frac{4s}{p-s+1}\right)^{s/2}\frac{1 - (4se^\eta/(p-s+1))^{s-\lfloor s/2\rfloor}}{1 - 4se^\eta/(p-s+1)}\\ &< 1 + (4s/(p-s+1))^{s/2}\frac{1}{1 - 4s/(p-s+1)\cdot e^\eta}.\end{aligned}$$

Here, (*) is due to $\eta > 0$ and $\lfloor s/2 \rfloor + 1 \geq s/2$. By setting $\eta = \log((p-s+1)/(8s))$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} \geq \frac{3sK}{4}\right) \\
=&\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2} \geq \frac{sK}{4}\right) \\
\leq&\frac{\mathbb{E}\exp\left(\eta(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2})\right)}{\exp\left(\eta \cdot \frac{sK}{4}\right)} \\
=&\frac{\prod_{k=1}^{K}\mathbb{E}\exp\left(\eta(w^{(k,m_1,m_2)} - \frac{s}{2})\right)}{\exp(\eta \cdot \frac{sK}{4})} \\
\leq&\left(1 + (4s/(p-s+1))^{s/2} \cdot 2\right)^K * \\
&\exp\left(-\frac{sK}{4}\log\left(\frac{p-s+1}{8s}\right)\right).
\end{aligned} \tag{A.17}$$

Since $p \geq 20s$ and $s \geq 4$, we have

$$\begin{aligned}
&\left(1 + 2(4s/(p-s+1))^{s/2}\right)^K \\
\leq&\exp\left(K\log\left(1 + 2\left(\frac{4}{p/s-1}\right)^{s/2}\right)\right) \\
\leq&\exp\left(K\log\left(1 + 2\left(\frac{4}{19}\right)^2\right)\right) \\
\leq&\exp\left(K \cdot 0.085\right) \leq \exp\left(sK\log(p/s) \cdot 0.0144\right), \\
&\exp\left(-\frac{sK}{4}\log\left(\frac{p-s+1}{8s}\right)\right) \\
=&\exp\left(-\frac{sK\log(p/s)}{4} + \frac{sK}{4}\log(8p/(p-s+1))\right) \\
\leq&\exp\left(-\frac{sK\log(p/s)}{4} + \frac{sK}{4}\log(8 \cdot 19/20)\right) \\
\leq&\exp\left(-sK\log(p/s) \cdot 0.08\right).
\end{aligned}$$

Combining the two inequalities above, we have

$$\begin{aligned}
&\left(1+(4s/(p-s+1))^{s/2}\cdot 2\right)^K \exp\left(-\frac{sK}{4}\log\left(\frac{p-s+1}{8s}\right)\right) \\
\leq&\exp\left(-c_0 sK\log(p/s)\right)
\end{aligned}$$

for $c_0 = 1/20$.

Next we choose $M = \lfloor \exp(c_0/2 \cdot sK\log(p/s)) \rfloor$. Note that

$$\begin{aligned}
&\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \\
=&\lambda \cdot \left(\left|\Omega^{(k,m_1)} \setminus \Omega^{(k,m_2)}\right| + \left|\Omega^{(k,m_2)} \setminus \Omega^{(k,m_1)}\right|\right) \\
=&\lambda\left(\left|\Omega^{(k,m_1)}\right| + \left|\Omega^{(k,m_2)}\right| - 2\left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|\right) \\
=&2\lambda\left(s - \left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|\right) \overset{(A.16)}{=} 2\lambda\left(s - w^{(k,m_1,m_2)}\right),
\end{aligned}$$

then we further have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{K}\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \geq \frac{sK\lambda}{2}, \forall 1 \leq m_1 < m_2 \leq M\right) \\
=\ &\mathbb{P}\left(\sum_{k=1}^{K} 2\lambda\left(s - w^{(k,m_1,m_2)}\right) \geq \frac{sK\lambda}{2}, \forall 1 \leq m_1 < m_2 \leq M\right) \\
=\ &\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} \leq \frac{3K}{4}, \forall 1 \leq m_1 < m_2 \leq M\right) \\
\overset{(A.17)}{\geq}& 1 - \frac{M(M-1)}{2}\exp\left(-c_0 sK\log(p/s)\right) \\
>\ & 1 - M^2\exp\left(-c_0 sK\log(p/s)\right) \geq 0,
\end{aligned}$$

which means there are positive probability that $\left\{\boldsymbol{\beta}^{(k,m)}\right\}_{\substack{k=1,\ldots,K\\m=1,\ldots,M}}$ satisfy

$$\begin{aligned}
\frac{sK\lambda}{2} &\leq \min_{1 \leq m_1 < m_2 \leq M} \sum_{k=1}^{K}\left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \\
&\leq \max_{1 \leq m_1 < m_2 \leq M} \sum_{k=1}^{K}\left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \leq 2sK\lambda.
\end{aligned} \tag{A.18}$$

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 (A.18).

Next, recall the canonical basis $\boldsymbol{e}_k = (0, \ldots, \overset{k\text{-th}}{\overbrace{1}}, 0, \cdots, 0) \in \mathbb{R}^p$. Define

$$\mathscr{T}^{(m)} = \sum_{k=1}^{K} \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}_k \circ \boldsymbol{e}_k, \quad 1 \leq m \leq M.$$

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:

$$\begin{aligned}
&D_{KL}\left((\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}), (\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\right) \\
:=&\mathbb{E}_{(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}\log\left(\frac{p(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{p(\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}\right).
\end{aligned}$$

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,

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}\left(\frac{p(\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{p(\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}\Big|\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}\right) \\
=&\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\left(\left(\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right)^\top \boldsymbol{u}_i\right) * \right. \\
&\left.\left(\boldsymbol{e}^{(k)\top}\boldsymbol{v}_i\right)\left(\boldsymbol{e}^{(k)\top}\boldsymbol{w}_i\right)\right)^2 \sigma^{-2}.
\end{aligned}$$

Therefore, for any $m_1 \neq m_2$,

$$\begin{aligned}
&D_{KL}\Big((\boldsymbol{y}^{(m_1)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}), (\boldsymbol{y}^{(m_2)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\Big)\\
=&\mathbb{E}_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^n \Big(\sum_{k=1}^K (\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)})^\top \boldsymbol{u}_i)\\
&(\boldsymbol{e}^{(k)\top}\boldsymbol{v}_i)(\boldsymbol{e}^{(k)\top}\boldsymbol{w}_i)\Big)^2 \sigma^{-2}\\
=&\frac{\sigma^{-2}}{2}\sum_{i=1}^n\sum_{k=1}^K \mathbb{E}_{\boldsymbol{u}}((\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)})^\top \boldsymbol{u}_i)^2\\
&\mathbb{E}_{\boldsymbol{v}}(\boldsymbol{e}^{(k)\top}\boldsymbol{v}_i)^2 \mathbb{E}_{\boldsymbol{w}}(\boldsymbol{e}^{(k)\top}\boldsymbol{w}_i)^2\\
=&\frac{n\sigma^{-2}}{2}\sum_{k=1}^K \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \leq \sigma^{-2} nKs\lambda.
\end{aligned}$$

Meanwhile, for any $1 \leq m_1 < m_2 \leq M$,

$$\begin{aligned}
&\|\mathscr{T}^{(m_1)} - \mathscr{T}^{(m_2)}\|_F\\
=&\left\|\sum_{k=1}^K (\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}) \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)}\right\|_F\\
=&\sqrt{\sum_{k=1}^K \left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2} \overset{\text{(A.18)}}{\geq} \sqrt{\frac{sK\lambda}{2}}.
\end{aligned}$$

By generalized Fano's Lemma (see, e.g., [60]),

$$\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T}\in\mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F \geq \sqrt{\frac{sK\lambda}{2}}\left(1 - \frac{\sigma^{-2} nKs\lambda + \log 2}{\log M}\right).$$

Finally we set $\lambda = \frac{c\sigma^2}{n}\log(p/s)$ for some small constant $c > 0$, then

$$\begin{aligned}
&\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T}\in\mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F^2 \geq \left(\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T}\in\mathcal{F}} \mathbb{E}\|\widehat{\mathscr{T}} - \mathscr{T}\|_F\right)^2\\
\geq&\frac{c\sigma^2\ sK\log(p/s)}{n}.
\end{aligned}$$

which has finished the proof of Theorem 6.

For the proof for Theorem 4, 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

$$\begin{aligned}
\frac{sK\lambda}{6} \leq& \min_{1\leq m_1<m_2\leq M} \sum_{k=1}^{K/3} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2\\
\leq& \max_{1\leq m_1<m_2\leq M} \sum_{k=1}^{K/3} \|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \leq \frac{2sK}{3}\lambda.
\end{aligned}$$

We then construct the following candidate symmetric tensors by blockwise design,

$$\mathcal{T}^{(m)} = \begin{cases}
\mathcal{T}^{(m)}_{[I_1,I_2,I_2]} = \sum_{k=1}^{K/3} \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)},\\
\mathcal{T}^{(m)}_{[I_2,I_1,I_2]} = \sum_{k=1}^{K/3} \boldsymbol{e}^{(k)} \circ \boldsymbol{\beta}^{(k,m)} \circ \boldsymbol{e}^{(k)},\\
\mathcal{T}^{(m)}_{[I_2,I_2,I_1]} = \sum_{k=1}^{K/3} \boldsymbol{e}^{(k)} \circ \boldsymbol{e}^{(k)} \circ \boldsymbol{\beta}^{(k,m)},\\
\mathcal{T}^{(m)}_{[I_1,I_1,I_1]}, \mathcal{T}^{(m)}_{[I_1,I_1,I_2]},\\
\mathcal{T}^{(m)}_{[I_1,I_2,I_1]}, \mathcal{T}^{(m)}_{[I_2,I_1,I_1]}, \mathcal{T}^{(m)}_{[I_2,I_2,I_2]} \text{ are all zeros.}
\end{cases}$$

Then we can see for any $\boldsymbol{u} \in \mathbb{R}^p$,

$$\langle \mathcal{T}^{(m)}, \boldsymbol{u}\circ\boldsymbol{u}\circ\boldsymbol{u}\rangle = 3\sum_{k=1}^{K/3}\Big(\boldsymbol{\beta}^{(k,m)\top}\boldsymbol{u}_{I_1}\Big)\cdot\Big(\boldsymbol{e}^{(k)\top}\boldsymbol{u}_{I_2}\Big)^2.$$

The rest of the proof essentially follows from the proof of Theorem 6. ∎

### *F. Proof of Theorem 7: High-Order Stein's Lemma*

The proof of this theorem follows from the one of Theorem 6 in [49]. For the sake of completeness, we restate the detail here. Applying the recursion representation of score function (A.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. ∎

### *G. Proofs of Lemmas 3, 4, and 5: Moment Calculation*

In this subsection, we present the detail proofs of moment calculation, including non-symmetric case, symmetric case, and interaction model.

*1) Proof of Lemma 3:* By the definition of $\{y_i\}$ in (VI.1) & (VI.2), we have

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{u}_i\circ\boldsymbol{v}_i\circ\boldsymbol{w}_i\Big) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{u}_i\circ\boldsymbol{v}_i\circ\boldsymbol{w}_i\Big)\\
&+\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \eta_k^*(\boldsymbol{\beta}_{1k}^{*\top}\boldsymbol{u}_i)(\boldsymbol{\beta}_{2k}^{*\top}\boldsymbol{v}_i)(\boldsymbol{\beta}_{3k}^{*\top}\boldsymbol{w}_i)\boldsymbol{u}_i\circ\boldsymbol{v}_i\circ\boldsymbol{w}_i\Big).
\end{aligned} \tag{A.19}$$

First, we observe $\mathbb{E}(\epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i) = 0$ due to the independence between $\epsilon_i$ 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

$$M = \mathbb{E}\Big((\boldsymbol{\beta}_1^{*\top} \boldsymbol{u})(\boldsymbol{\beta}_2^{*\top} \boldsymbol{v})(\boldsymbol{\beta}_3^{*\top} \boldsymbol{w})\boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\Big) \in \mathbb{R}^{p_1 \times p_2 \times p_3}. \tag{A.20}$$

Each entry of $M$ can be calculated as follows

$$\begin{aligned} M_{ijk} &= \mathbb{E}\Big((\boldsymbol{\beta}_1^{*\top} \boldsymbol{u})(\boldsymbol{\beta}_2^{*\top} \boldsymbol{v})(\boldsymbol{\beta}_3^{*\top} \boldsymbol{w}) u_i v_j w_k\Big) \\ &= \mathbb{E}\Big((\beta_{1i}^* u_i + \sum_{m \neq i} \beta_{1m}^* u_m) u_i\Big) \\ &\quad \times \mathbb{E}\Big((\beta_{2j}^* u_i + \sum_{m \neq j} \beta_{2m}^* v_m) v_j\Big) \\ &\quad \times \mathbb{E}\Big((\beta_{3k}^* w_k + \sum_{m \neq k} \beta_{3m}^* w_m) w_k\Big) \\ &= \beta_{1i}^* \beta_{2j}^* \beta_{3k}^*, \end{aligned}$$

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. ■

*2) 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 (A.20) 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

$$\begin{aligned} M_{s_{ijk}} &= \mathbb{E}\big(\beta_i^* x_i + \beta_j^* x_j + \beta_k^* x_k\big)^3 x_i x_j x_k \\ &= 6\beta_i^* \beta_j^* \beta_k^*. \end{aligned}$$

- 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

$$\begin{aligned} & \mathbb{E}\Big(M_s - \sum_{m=1}^p \Big(M_1 \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m \\ & \quad + \boldsymbol{e}_m \circ M_1 \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ M_1\Big)\Big) \\ = & \ 6\sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*. \end{aligned}$$

This finishes our proof. ■

*3) 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. ■

### H. Proof of Lemma 6

Recall the $\|X\|_{\psi_\alpha}$ is defined in Definition 1. 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 [61]), we have

$$\Big\|\sum_{i=1}^n a_iX_i\Big\|_p \leq 2\Big\|\sum_{i=1}^n a_i\varepsilon_iX_i\Big\|_p = 2\Big\|\sum_{i=1}^n a_i\varepsilon_i|X_i|\Big\|_p, \tag{A.21}$$

where $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables and we notice that $\varepsilon_iX_i$ 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

$$2\Big\|\sum_{i=1}^n a_i\varepsilon_i|X_i|\Big\|_p \leq 2\Big\|\sum_{i=1}^n a_i\varepsilon_i(\beta + Z_i)\Big\|_p. \tag{A.22}$$

By triangle inequality,

$$\begin{aligned} &2\Big\|\sum_{i=1}^n a_i\varepsilon_i(\beta + Z_i)\Big\|_p \\ \leq\ &2\Big\|\sum_{i=1}^n a_i\varepsilon_iZ_i\Big\|_p + 2\Big\|\sum_{i=1}^n a_i\varepsilon_i\beta\Big\|_p. \end{aligned} \tag{A.23}$$

Next, we will bound the second term of the RHS of (A.23). 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

$$\begin{aligned} \Big\|\sum_{i=1}^n a_i\varepsilon_i\beta\Big\|_p &\leq \Big(\frac{p-1}{2-1}\Big)^{1/2}\Big\|\sum_{i=1}^n a_i\varepsilon_i\beta\Big\|_2 \\ &\leq \beta\sqrt{p}\Big\|\sum_{i=1}^n a_i\varepsilon_i\Big\|_2. \end{aligned} \tag{A.24}$$

Since $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables, some simple calculations implies

$$\Big(\mathbb{E}\Big(\sum_{i=1}^n \varepsilon_ia_i\Big)^2\Big)^{1/2} \tag{A.25}$$

$$\begin{aligned} &= \Big(\mathbb{E}\Big(\sum_{i=1}^n \varepsilon_i^2\ a_i^2 + 2\sum_{1\leq i<j\leq n}\varepsilon_i\varepsilon_ja_ia_j\Big)\Big)^{1/2} \\ &= \Big(\sum_{i=1}^n a_i^2\mathbb{E}\varepsilon_i^2 + 2\sum_{1\leq i<j\leq n}a_ia_j\mathbb{E}\varepsilon_i\mathbb{E}\varepsilon_j\Big)^{1/2} \\ &= \Big(\sum_{i=1}^n a_i^2\Big)^{1/2} = \|\boldsymbol{a}\|_2. \end{aligned} \tag{A.26}$$

Combining inequalities (A.22)-(A.25),

$$2\Big\|\sum_{i=1}^n a_i\varepsilon_i|X_i|\Big\|_p \leq 2\Big\|\sum_{i=1}^n a_i\varepsilon_iZ_i\Big\|_p + 2\beta\sqrt{p}\|\boldsymbol{a}\|_2. \tag{A.27}$$

Let $\{Y_i\}_{i=1}^n$ are independent symmetric random variables satisfying $\mathbb{P}(|Y_i| \geq t) = \exp(-t^\alpha)$ for all $t \geq 0$. Then we have

$$\begin{aligned} &\mathbb{P}(Z_i \geq t) \\ \leq&\mathbb{P}(|X_i| \geq t+\beta) = \mathbb{P}\left(\exp(|X_i|^\alpha) \geq \exp((t+\beta)^\alpha)\right) \\ \leq&\mathbb{E}(\exp(|X_i|)^\alpha)\cdot\exp(-(t+\beta)^\alpha) \leq 2\exp(-(t+\beta)^\alpha) \\ \leq&2\exp(-t^\alpha - \beta^\alpha) = \mathbb{P}(|Y_i| \geq t), \end{aligned}$$

which implies

$$\Big\|\sum_{i=1}^n a_i\varepsilon_iZ_i\Big\|_p \leq \Big\|\sum_{i=1}^n a_i\varepsilon_iY_i\Big\|_p = \Big\|\sum_{i=1}^n a_iY_i\Big\|_p, \tag{A.28}$$

since $\varepsilon_iY_i$ and $Y_i$ have the same distribution due to symmetry. Combining (A.27) and (A.28) together, we reach

$$\Big\|\sum_{i=1}^n a_iX_i\Big\|_p \leq 2\beta\sqrt{p}\|\boldsymbol{a}\|_2 + 2\Big\|\sum_{i=1}^n a_iY_i\Big\|_p. \tag{A.29}$$

For $0 < \alpha < 1$, it follows Lemma 25 that

$$\Big\|\sum_{i=1}^n a_iY_i\Big\|_p \leq C_1(\alpha)(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty), \tag{A.30}$$

where $C_1(\alpha)$ is some absolute constant only depending on $\alpha$.

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 $X = |\sum_{i=1}^n a_iY_i|^p$ and changing the variable $t = t^p$, then we have

$$\begin{aligned} &\mathbb{E}|\sum_{i=1}^n a_iY_i|^p \\ =&\int_0^\infty \mathbb{P}\Big(|\sum_{i=1}^n a_iY_i| \geq t\Big)pt^{p-1}dt \\ \leq&\int_0^\infty 2\exp\Big(-c\min\Big(\frac{t^2}{\|\boldsymbol{a}\|_2^2}, \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big)pt^{p-1}dt, \end{aligned} \tag{A.31}$$

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 $\frac{t^2}{\|\boldsymbol{a}\|_2^2} \leq \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (A.31) 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,

$$\Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \frac{p^{1/p}}{(2c)^{1/2}}\sqrt{p}\|\boldsymbol{a}\|_2 \leq \frac{\sqrt{2}}{\sqrt{c}}\sqrt{p}\|\boldsymbol{a}\|_2. \tag{A.32}$$

2) If $\frac{t^2}{\|\boldsymbol{a}\|_2^2} > \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (A.31) 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,

$$\begin{aligned} \Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} &\leq \frac{2p^{1/p}}{(c\alpha)^{1/\alpha}}p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*} \\ &\leq \frac{4}{(c\alpha)^{1/\alpha}}p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}. \end{aligned} \tag{A.33}$$

3) Overall, we have the following by combining (A.32) and (A.33),

$$\begin{aligned} & \Big(\mathbb{E}|\textstyle\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \\ & \leq \max\Big(\sqrt{\tfrac{2}{c}}, \tfrac{4}{(c\alpha)^{1/\alpha}}\Big)\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\Big). \end{aligned}$$

After denoting $C_2(\alpha) = \max\Big(\sqrt{\tfrac{2}{c}}, \tfrac{4}{(c\alpha)^{1/\alpha}}\Big)$, we reach

$$\Big\|\sum_{i=1}^n a_i Y_i\Big\|_p \leq C_2(\alpha)\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\Big). \tag{A.34}$$

Since $0 < \beta < 1$, the conclusion can be reached by combining (A.29),(A.30) and (A.34). ■

### I. 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,

$$\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2, \tag{A.35}$$

where

$$\begin{aligned} \mathcal{E}_1 =& \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^* \rangle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \\ & - \sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*. \\ \mathcal{E}_2 =& \frac{1}{n}\sum_{i=1}^n \epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i. \end{aligned}$$

**Bounding $\mathcal{E}_1$:** For $k$-th componet of $\mathcal{E}_1$, we denote

$$\begin{aligned} \mathcal{E}_{1k} =& \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i, \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^* \rangle \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \\ & - \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*. \end{aligned}$$

Define

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

By using Lemma 2 and $s \leq d \leq Cs$, it suffices to have for some absolute constant $C_{11}$,

$$\|\mathcal{E}_{1k}\|_{s+d} \leq C_{11}\delta_{n,p,s},$$

with probability at least $1 - 10/n^3$, where $\|\cdot\|_{s+d}$ is the sparse tensor spectral norm defined in (II.3). Equipped with the triangle inequality, the sparse tensor spectral norm for $\mathcal{E}_1$ can be bounded by

$$\|\mathcal{E}_1\|_{s+d} \leq C_{11}\delta_{n,p,s}\sum_{k=1}^K \eta_k^*, \tag{A.36}$$

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 $C_{12}$

$$\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i\Big\|_{s+d} \leq C_{12}\|\boldsymbol{\epsilon}\|_\infty C_{11}\delta_{n,p,s},$$

with probability at least $1 - 1/p$. According to Lemma 23, we have

$$\mathbb{P}\Big(\|\mathcal{E}_2\|_{s+d} \geq C_{12}\sigma\log n\delta_{n,p,s}\Big) \leq \frac{1}{p} + \frac{3}{n} \leq \frac{4}{n}. \tag{A.37}$$

**Bounding $\mathcal{E}$:** Putting (A.36) and (A.37) together, we obtain

$$\|\mathcal{E}\|_{s+d} \leq \Big(C_{11}\sum_{k=1}^K \eta_k^* + C_{12}\sigma\log n\Big)\delta_{n,p,s},$$

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. ■

### J. 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

$$\widetilde{\boldsymbol{\beta}}_k^{(t+1)} = \varphi_{\frac{\mu}{\phi} h(\boldsymbol{\beta}_k^{(t)})}\Big(\boldsymbol{\beta}_k^{(t)} - \frac{\mu}{\phi}\nabla_k \mathcal{L}(\boldsymbol{\beta}_k^{(t)})_{F^{(t)}}\Big). \quad \text{(A.38)}$$

Here,

- $h(\boldsymbol{\beta}_k^{(t)})$ is the $k$-th component of $h(\boldsymbol{B}^{(t)})$ defined in (III.10).
- $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}) = (\nabla_1 \mathcal{L}(\boldsymbol{\beta}_1), \cdots, \nabla_K \mathcal{L}(\boldsymbol{\beta}_K))$.
- $F^{(t)} = \cup_{k=1}^K F_k^{(t)}$, where $F_k^{(t)} = \text{supp}(\boldsymbol{\beta}_k^*) \cup \text{supp}(\boldsymbol{\beta}_k^{(t)})$.
- For a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a subset $A \subset \{1,\dots,p\}$, 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 (A.13) is satisfied and $|F| \lesssim Ks$. As long as the step size $\mu \leq 32\ R^{-20/3}/(3K[220+270K]^2)$, we obtain the upper bound for $\{\widetilde{\boldsymbol{\beta}}_k^+\}$,

$$\begin{aligned}&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\ \leq&\Big(1 - 32\mu\frac{R^{-\frac{8}{3}}}{K^2}\Big)\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\ &+2C_3\mu^2 R^{-\frac{8}{3}} \eta_{\min}^{*-\frac{4}{3}} \frac{\sigma^2 K^{-2} s \log p}{n}, \end{aligned} \quad \text{(A.39)}$$

with probability at least $1-(21K^2+11K+4Ks)/n$.

The proof of Lemma 13 is postponed to the Section L. 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

$$\begin{aligned}h(\boldsymbol{\beta}_k) =&\frac{\sqrt{4\log np}}{n} * \\ &\sqrt{\sum_{i=1}^n \Big(\sum_{k=1}^K \eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3 - y_i\Big)^2 \Big(\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2\Big)^2}.\end{aligned} \quad \text{(A.40)}$$

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 L. By using Lemma 14 and induction, we have

$$F^{(t+1)} \subset \cdots F^{(1)} \subset F^{(0)} = \cup_{k=1}^K \text{supp}(\boldsymbol{\beta}_k^*) \cup \text{supp}(\boldsymbol{\beta}_k^{(0)}).$$

It implies for every $t$, we have $|F^{(t)}| \lesssim Ks$. Combining with Lemmas 13 and 14 together, we obtain with probability at least $1-(21K^2+11K+4Ks)/n$,

$$\begin{aligned}&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\ \leq&\Big(1 - 32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\ &+2C_3\mu^2 R^{-\frac{8}{3}} \eta_{\min}^{*-\frac{4}{3}} \frac{\sigma^2 K^{-2} s \log p}{n}, \end{aligned} \quad \text{(A.41)}$$

This ends the proof. ■

### K. 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

$$\begin{aligned}&\Big\|\mathscr{T} - \mathscr{T}^*\Big\|_F^2 \\ =&\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K \eta_k \beta_{ki_1}\beta_{ki_2}\beta_{ki_3} - \sum_{k=1}^K \eta_k^*\beta_{ki_1}^*\beta_{ki_2}^*\beta_{ki_3}^*\Big)^2.\end{aligned}$$

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

$$\begin{aligned}&\Big\|\mathscr{T} - \mathscr{T}^*\Big\|_F^2 \\ =&\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K \bar{\beta}_{ki_1}\bar{\beta}_{ki_2}\bar{\beta}_{ki_3} - \sum_{k=1}^K \bar{\beta}_{ki_1}^*\bar{\beta}_{ki_2}^*\bar{\beta}_{ki_3}^*\Big)^2 \\ =&\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K (\bar{\beta}_{ki_1} - \bar{\beta}_{ki_1}^*)\bar{\beta}_{ki_2}^*\bar{\beta}_{ki_3}^* \\ &+\sum_{k=1}^K \bar{\beta}_{ki_1}(\bar{\beta}_{ki_2} - \bar{\beta}_{ki_2}^*)\bar{\beta}_{ki_3}^* \\ &+\sum_{k=1}^K \bar{\beta}_{ki_1}\bar{\beta}_{ki_2}(\bar{\beta}_{ki_3} - \bar{\beta}_{ki_3}^*)\Big)^2 = \text{RHS}.\end{aligned}$$

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

$$\begin{aligned}\text{RHS} \le & 3\sum_{i_1,i_2,i_3}\Big[\sum_{k=1}^K(\bar\beta_{ki_1}-\bar\beta^*_{ki_1})^2\sum_{k=1}^K\bar\beta^{*2}_{ki_2}\bar\beta^{*2}_{ki_3}\\ &+\sum_{k=1}^K(\bar\beta_{ki_2}-\bar\beta^*_{ki_2})^2\sum_{k=1}^K\bar\beta^2_{ki_1}\bar\beta^{*2}_{ki_3}\\ &+\sum_{k=1}^K(\bar\beta_{ki_3}-\bar\beta^*_{ki_3})^2\sum_{k=1}^K\bar\beta^2_{ki_2}\ \bar\beta^2_{ki_1}\Big]\end{aligned}$$

At the same time, using $\eta_k \le (1+c)\eta_k^*$ for $k\in[K]$,

$$\begin{aligned}&\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2\\ \le\ & 3\Big[\sum_{i_1=1}^p\sum_{k=1}^K(\bar\beta_{ki_1}-\bar\beta^*_{ki_1})^2(\sum_{i_2=1}^p\sum_{i_3=1}^p\sum_{k=1}^K\bar\beta^{*2}_{ki_2}\bar\beta^{*2}_{ki_3})\\ &+\sum_{i_2=1}^p\sum_{k=1}^K(\bar\beta_{ki_2}-\bar\beta^*_{ki_2})^2(\sum_{i_1=1}^p\sum_{i_3=1}^p\sum_{k=1}^K\bar\beta^{2}_{ki_1}\bar\beta^{*2}_{ki_3})\\ &+\sum_{i_3=1}^p\sum_{k=1}^K(\bar\beta_{ki_3}-\bar\beta^*_{ki_3})^2(\sum_{i_2=1}^p\sum_{i_1=1}^p\sum_{k=1}^K\bar\beta^{2}_{ki_2}\bar\beta^{2}_{ki_1})\Big]\\ =\ & 3\Big(\sum_{k=1}^K\|\bar{\boldsymbol\beta}_k-\bar{\boldsymbol\beta}^*_k\|_2^2\Big)\Big(\sum_{k=1}^K(\sqrt[3]{\eta_k^*})^4\\ &+\sum_{k=1}^K(\sqrt[3]{\eta_k^*})^2(\sqrt[3]{\eta_k})^2+\sum_{k=1}^K(\sqrt[3]{\eta_k})^4\Big)\\ \le\ & 9(1+c)\Big(\sum_{k=1}^K\|\bar{\boldsymbol\beta}_k-\bar{\boldsymbol\beta}^*_k\|_2^2\Big)\Big(\sum_{k=1}^K(\sqrt[3]{\eta_k^*})^4\Big).\end{aligned}$$

For the non-symmetric tensor estimation model, we have

$$\begin{aligned}&\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2\\ &=\sum_{i_1,i_2,i_3}\Big(\sum_{k=1}^K\eta_k\beta_{1ki_1}\beta_{2ki_2}\beta_{3ki_3}-\sum_{k=1}^K\eta_k^*\beta^*_{1ki_1}\beta^*_{2ki_2}\beta^*_{3ki_3}\Big)^2.\end{aligned}$$

Following the same strategy above, we obtain

$$\begin{aligned}&\Big\|\mathscr{T}-\mathscr{T}^*\Big\|_F^2\\ &\le 3(1+c)\Big(\sum_{k=1}^K\|\bar{\boldsymbol\beta}_{1k}-\bar{\boldsymbol\beta}^*_{1k}\|_2^2+\sum_{k=1}^K\|\bar{\boldsymbol\beta}_{2k}-\bar{\boldsymbol\beta}^*_{2k}\|_2^2\\ &\quad+\sum_{k=1}^K\|\bar{\boldsymbol\beta}_{3k}-\bar{\boldsymbol\beta}^*_{3k}\|_2^2\Big)\Big(\sum_{k=1}^K(\sqrt[3]{\eta_k^*})^4\Big).\end{aligned}$$

This ends the proof. ■

### L. Proof of Lemma 13

First of all, we state a lemma to illustrate the effect of weight $\phi$. The proof of Lemma 15 is deferred to Section 15.

*Lemma 15:* Consider $\{y_i\}_{i=1}^n$ come from either non-symmetric tensor estimation model (VI.1) or symmetric tensor estimation model (III.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 $\Gamma$ is the incoherence parameter defined in Definition 3.

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 (A.13), $\max_k|\eta_k-\eta_k^*|\le\varepsilon_0$ for some small $\varepsilon_0$. Based on those two facts described above, we replace $\eta_k$ by $\eta_k^*$ and $\phi$ 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 [62]. Therefore, we define the weighted estimator and weighted true parameter as $\bar{\boldsymbol\beta}_k=\sqrt[3]{\eta_k^*}\boldsymbol\beta_k$, $\bar{\boldsymbol\beta}_k^*=\sqrt[3]{\eta_k^*}\boldsymbol\beta_k^*$. Now, $\eta_k^*\beta_k\circ\beta_k\circ\beta_k=\bar\beta_k\circ\bar\beta_k\circ\bar\beta_k$. Recall $\cdot$ is the loss function defined in (III.4). Correspondingly with a slight abuse of notation, 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

$$\begin{aligned}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol\beta}_k)_F=&\frac{6\sqrt[3]{\eta_k^*}}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K(\boldsymbol{x}_{i_F}^\top\bar{\boldsymbol\beta}_{k'})^3\\ &-\sum_{k'=1}^K(\boldsymbol{x}_{i_F}^\top\bar{\boldsymbol\beta}^*_{k'})^3\Big)(\boldsymbol{x}_{i_F}^\top\bar{\boldsymbol\beta}_k)^2\boldsymbol{x}_{i_F}.\end{aligned}\tag{A.42}$$

According to the definition of thresholding function (III.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

$$\begin{aligned}h(\bar{\boldsymbol\beta}_k)=&\frac{\sqrt{4\log(np)}}{n}*\\ &\sqrt{\sum_{i=1}^n\Big(\sum_{k=1}^K(\boldsymbol{x}_{i_F}^\top\bar{\boldsymbol\beta}_k)^3-y_i\Big)^2\eta_k^{*\frac{2}{3}}(\boldsymbol{x}_{i_F}^\top\bar{\boldsymbol\beta}_k)^2}.\end{aligned}\tag{A.43}$$

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:

$$\begin{aligned}
&\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k^*}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2\\
&=\sum_{k=1}^{K}\Big\|\boldsymbol{z}_k - \frac{\mu\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k) + \frac{\mu\sqrt[3]{\eta_k^*}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\gamma}_k\Big\|_2^2\\
&=\underbrace{\sum_{k=1}^{K}\Big\|\boldsymbol{z}_k - \mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2}_{\text{A: gradient update effect}}\\
&\quad+\underbrace{\sum_{k=1}^{K}\Big\|\frac{\mu\sqrt[3]{\eta_k^*}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\gamma_k\Big\|_2^2}_{\text{B: threshoding effect}}\\
&\quad+\underbrace{\sum_{k=1}^{K}\Big\langle\boldsymbol{z}_k - \mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k), \frac{\mu\sqrt[3]{\eta_k}}{\phi}h(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\gamma}_k\Big\rangle}_{\text{C: cross term}}. \quad (A.44)
\end{aligned}$$

In the following proof, we will bound three parts sequentially.

*1) 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

$$\begin{aligned}
A =&\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2 - 2\mu\sum_{k=1}^{K}\Big\langle\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{z}_k\Big\rangle\\
&+\mu^2\sum_{k=1}^{K}\Big\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2\\
\le\ &\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2 - 2\mu\underbrace{\sum_{k=1}^{K}\Big\langle\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{z}_k\Big\rangle}_{A_1}\\
&+2\mu^2\underbrace{\sum_{k=1}^{K}\Big\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2}_{A_2}\\
&+2\mu^2\underbrace{\sum_{k=1}^{K}\Big\|\frac{\sqrt[3]{\eta_k^*}}{\phi}\Big(\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k) - \nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big)\Big\|_2^2}_{A_3}\\
&+2\mu\underbrace{\sum_{k=1}^{K}\Big\langle\boldsymbol{z}_k, \frac{\sqrt[3]{\eta_k^*}}{\phi}\Big(\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k) - \nabla_k\mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big)\Big\rangle}_{A_4}, \quad (A.45)
\end{aligned}$$

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 [52].

**Step One: Lower bound for $A_1$.** Plugging in $\phi = (\sum_{k=1}^{K}\eta_k^*)^2$, we have

$$K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}} \le \frac{(\sqrt[3]{\eta_k^*})^2}{\phi} = \frac{(\sqrt[3]{\eta_k^*})^2}{(\sum_{k=1}^{K}\eta_k^*)^2} \le K^{-2}R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}. \quad (A.46)$$

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 \ge& 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} \quad (A.47)\\
&+\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}, \quad (A.48)
\end{aligned}$$

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 A.2: 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 A.2, the expectation of $A_{11}$ can be calculated explicitly as

$$\begin{aligned}\mathbb{E}(A_{11}) = 36&\sum_{k=1}^K\sum_{k'=1}^K(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)^2(\boldsymbol{z}_{k'}^\top\boldsymbol{z}_k) \Leftarrow\ I_1\\ &+72\sum_{k=1}^K\sum_{k'=1}^K(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'}^*) \Leftarrow\ I_2\\ &+108\sum_{k=1}^K\sum_{k'=1}^K(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_k^*) \Leftarrow\ I_3\\ &+54\sum_{k=1}^K\sum_{k'=1}^K(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k'}^*)(\bar{\boldsymbol{\beta}}_k^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\boldsymbol{z}_k) \Leftarrow\ I_4.\end{aligned} \tag{A.49}$$

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 $\Gamma$ is the incoherence parameter. Putting the above four bounds together, they jointly provide

$$\begin{aligned}\mathbb{E}(A_{11})\geq&36\eta_{\min}^{*4/3}\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\\ &-\Big(36\eta_{\min}^{*4/3}\Gamma^2+180\eta_{\min}^{*4/3}\Gamma\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2.\end{aligned} \tag{A.50}$$

On the other hand, repeatedly using Lemma 1, we obtain that with probability at least $1-1/n$,

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^n\Big((\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\\ &-\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big)\Big|\\ &\leq C\frac{(\log n)^3}{\sqrt{n}}(\sqrt[3]{\eta_{\max}^*})^4\|\boldsymbol{z}_{k'}\|_2\|\boldsymbol{z}_k\|_2.\end{aligned}$$

Taking the summation over $k, k'\in[K]$, it could further imply that for some absolute constant $C$,

$$\Big|A_{11}-\mathbb{E}(A_{11})\Big|\leq 18C\frac{(\log n)^3}{\sqrt{n}}(\sqrt[3]{\eta_{\max}^*})^4\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2, \tag{A.51}$$

with probability at least $1-K^2/n$. Combining (A.50) and (A.51), we obtain with probability at least $1-K^2/n$,

$$\begin{aligned}K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}A_{11}\geq\Big[&36K^{-2}R^{-\frac{8}{3}}-K^{-\frac{3}{2}}\Big(216R^{-\frac{8}{3}}\Gamma\\ &+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big]\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2,\end{aligned} \tag{A.52}$$

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 A.2, 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 (A.13), $\|\boldsymbol{z}_k\|_2\leq\varepsilon_0\sqrt[3]{\eta_k^*}$, it suffices to show that

$$\begin{aligned}&\sum_{k=1}^K\sum_{k'=1}^K(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'})^2(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*)\\ \geq\ &-\sum_{k=1}^K\sum_{k'=1}^K\|\boldsymbol{z}_k\|_2^2\|\bar{\boldsymbol{\beta}}_{k'}^*\|_2^2\|\boldsymbol{z}_k'\|_2\|\bar{\boldsymbol{\beta}}_k^*\|_2\\ \geq\ &-\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2.\end{aligned}$$

This immediately implies a lower bound for $\mathbb{E}(A_{12})$ after we bound similarly for $I_2, I_3$ and $I_4$,

$$\mathbb{E}(A_{12})\geq -270\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2. \tag{A.53}$$

By Lemma 1, we obtain for some absolute constant $C$,

$$\begin{aligned}&K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}A_{12}\\ \geq& K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}\Big[\mathbb{E}(A_{12})-18C\eta_{\max}^{*\frac{4}{3}}\varepsilon_0\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2\Big)^2\frac{(\log n)^3}{\sqrt{n}}\Big]\\ \geq& -K^{-1}R^{-\frac{2}{3}}\varepsilon_0\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big),\end{aligned}\tag{A.54}$$

with probability at least $1-K^2/n$. The detail derivation is the same as in (A.52), so we omit here.

Similarly, the lower bounds of $A_{13}$ to $A_{19}$ can be derived as follows

$$\begin{aligned}&K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{14}\\ \geq& -K^{\frac{1}{2}}\varepsilon_0\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{13},A_{15},A_{17}\\ \geq& -K^{\frac{1}{2}}\varepsilon_0^2\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{16},A_{18}\\ \geq& -K^{\frac{1}{2}}\varepsilon_0^3\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big)\\ &K^{-\frac{1}{2}}\eta_{\max}^{*-\frac{4}{3}}A_{19}\\ \geq& -K^{\frac{1}{2}}\varepsilon_0^4\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big).\end{aligned}\tag{A.55}$$

Putting (A.52), (A.54) and (A.55) together, we have with probability at least $1-9K^2/n$,

$$\begin{aligned}A_1 \;\geq\; &\Big[36K^{-2}R^{-\frac{8}{3}}-K^{-\frac{3}{2}}\Big(2160R^{-\frac{3}{3}}\Gamma+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\\ &-8\varepsilon_0K^{-1}R^{-\frac{2}{3}}\Big(270+18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big]\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big).\end{aligned}$$

For the above bound,

- When the sample size satisfies
$$n\geq(18CK^{1/2}R^{8/3}(\log n)^3)^2,$$
we have
$$\begin{aligned}&\max\Big\{18K^{-\frac{3}{2}}C\frac{(\log n)^3}{\sqrt{n}},8\varepsilon_0\;K^{-1}R^{-\frac{2}{3}}18C\frac{(\log n)^3}{\sqrt{n}}\Big\}\\ \leq& K^{-2}R^{-\frac{8}{3}}.\end{aligned}$$
- 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 (A.13). Thus, we are able to simplify $A_1$ by

$$A_1\geq 32K^{-2}R^{-\frac{8}{3}}\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big),\tag{A.56}$$

with probability at least $1-9K^2/n$.

**Step Two: Upper bound for $A_2$.** We observe the fact that

$$\begin{aligned}A_2=&\sum_{k=1}^K\Big\|\frac{1}{\phi}\sqrt[3]{\eta_k^*}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2\\ =&\sup_{\boldsymbol{w}\in\mathbb{S}^{Ks-1}}\Big|\Big\langle\sum_{k=1}^K\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k),\boldsymbol{w}\Big\rangle\Big|^2,\end{aligned}\tag{A.57}$$

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 (A.42), $A_2'$ is explicitly written as

$$\begin{aligned}A_2'=&\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^3-\sum_{k'=1}^K(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'}^*)^3\Big)*\\ &\Big(\sum_{k=1}^K\frac{(\sqrt[3]{\eta_k^*})^2}{\phi}(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2(\boldsymbol{x}_i^\top\boldsymbol{w})\Big).\end{aligned}$$

Following by (A.46) and (A.48), 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 $\boldsymbol{x}_i^\top\boldsymbol{w}$.

$$\begin{aligned}A_2'\leq K^{-2}R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}\Big[&\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\\ &\sum_{k=1}^K(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big)\Leftarrow A_{21}'\\ &+\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\\ &\sum_{k=1}^K2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big)\Leftarrow A_{22}'\\ &+\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\\ &\sum_{k=1}^K(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\boldsymbol{w})\Big)\Leftarrow A_{23}'\\ &+\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\\ &\sum_{k=1}^K(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big)\Leftarrow A_{24}'\\ &+\frac{6}{n}\sum_{i=1}^n\Big(\sum_{k'=1}^K3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\\ &\sum_{k=1}^K2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big)\Leftarrow A_{25}'\end{aligned}$$

$$
\begin{aligned}
&+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\\
&\quad\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big)\Leftarrow\ A'_{26}\\
&+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\\
&\quad\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big)\Leftarrow\ A'_{27}\\
&+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\\
&\quad\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big)\Leftarrow\ A'_{28}\\
&+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\\
&\quad\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\boldsymbol{w})\Big)\Big].\Leftarrow\ A'_{29}
\end{aligned}
$$

Let's bound $A'_{21}$ first. By using the same technique when calculating $\mathbb{E}(A_{11})$ in (A.49), 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 (II.3), it suffices to bound $A'_{21}$ by

$$
\begin{aligned}
R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}K^{-\frac{1}{2}}A'_{21}\leq K^{-2}R^2\Big[&216+54K\\
&+216K\Gamma+18CK\delta_{n,p,s}\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big)
\end{aligned}
$$

with probability at least $1-10K^2/n^3$, where $\delta_{n,p,s}$ is defined in (IV.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

$$
A_2\leq K^{-1}R^4[220+270K]^2\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big),\tag{A.58}
$$

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 (A.46). For fixed $\{\epsilon_i\}_{i=1}^n$, applying Lemma 1, we have

$$
\begin{aligned}
&\Big|\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}-\mathbb{E}\Big(\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}\Big)\Big|\\
\leq\ & C(\log n)^{\frac{3}{2}}\|\boldsymbol{\epsilon}\|_2\|\bar{\boldsymbol{\beta}}_k\|_2^2,
\end{aligned}
$$

with probability at least $1-1/n$. Together with Lemma 23, we obtain for any $j\in[Ks]$,

$$
\Big|\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}\Big|\leq 6CC_0\sigma\|\bar{\boldsymbol{\beta}}_k\|_2^2\frac{(\log n)^{3/2}}{\sqrt{n}},
$$

with probability at least $1-4/n$, where $\sigma$ is the noise level. According to (A.13),

$$
\Big\|\bar{\boldsymbol{\beta}}_k-\bar{\boldsymbol{\beta}}_k^*\Big\|_2^2\leq\sum_{k=1}^{K}\Big\|\bar{\boldsymbol{\beta}}_k-\bar{\boldsymbol{\beta}}_k^*\Big\|_2^2\leq K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2,
$$

which further implies $\|\bar{\boldsymbol{\beta}}_k\|_2^2\leq(1+K^{\frac{1}{2}}\varepsilon_0)^2\eta_{\max}^{*\frac{2}{3}}$. Equipped with union bound over $j\in[Ks]$,

$$
\begin{aligned}
&\max_{j\in[Ks]}\Big|\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}\Big|\\
\leq\ & 6CC_0\sigma(1+K^{\frac{1}{2}}\varepsilon_0)^2(\sqrt[3]{\eta_{\max}^*})^2\frac{(\log n)^{3/2}}{\sqrt{n}},
\end{aligned}
$$

with probability at least $1-4Ks/n$. Letting $C=6C_0(Ce)^{-2/3}(1+K^{\frac{1}{2}}\varepsilon_0)^2$,

$$
A_3\leq C\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}\sigma^2K^{-2}\frac{s(\log n)^3}{n},\tag{A.59}
$$

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 $1-4K/n$,

$$
A_4\leq 2C\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}K^{-1}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{2}{3}}.\tag{A.60}
$$

This term is negligible in terms of the order when comparing with (A.59).

**Summary.** Putting the bounds (A.56), (A.58), (A.59) and (A.60) together, we achieve an upper bound for gradient update effect as follows,

$$\begin{aligned} A \leq & \Big(1 - 64\mu K^{-2} R^{-\frac{8}{3}} \\ + & \quad 2\mu^2 K^{-1} R^4 [220 + 270K]^2\Big) \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2 \\ & + 4\mu C K^{-2} \eta_{\min}^{*-\frac{4}{3}} R^{\frac{8}{3}} \frac{\sigma^2 s (\log n)^3}{n}, \end{aligned} \tag{A.61}$$

with probability at least $1 - (18K^2 + 4K + 4Ks)/n$. ■

*2) Bounding Thresholding Effect:* The thresholding effect term in (A.44) can also be decomposed into optimization error and statistical error. Recall that $B$ can be explicitly written as

$$\begin{aligned} B \;=\; & \sum_{k=1}^{K} \Big\| \mu \frac{\eta_k^{*\frac{2}{3}}}{\phi} \frac{4\sqrt{\log(np)}}{n} * \\ & \sqrt{\sum_{i=1}^{n} \Big( \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 - y_i \Big)^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \gamma_k} \Big\|_2^2, \end{aligned}$$

where $\text{supp}(\gamma_k) \subset F_k$ and $\|\gamma_k\|_\infty \leq 1$. By using $(a+b)^2 \leq 2(a^2 + b^2)$, we have

$$B \leq \mu^2 \frac{64Ks\log p}{n} (B_1 + B_2),$$

where

$$\begin{aligned} B_1 = & \frac{1}{n} \sum_{i=1}^{n} \Big( \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 \\ & - \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3 \Big) \Big( \sum_{k=1}^{K} \frac{\eta_k^{*\frac{4}{3}}}{\phi^2} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \Big) \\ B_2 = & \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \sum_{k=1}^{K} \frac{\eta_k^{*\frac{4}{3}}}{\phi^2} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4. \end{aligned}$$

**Bounding** $B_1$**.** This optimization error term shares similar structure with (A.57) but with higher order. Therefore, we follow the same idea as we did in bounding (A.57). Following by (A.46) and some basic expansions and inequalities,

$$\begin{aligned} B_1 \leq & K^{-2} R^{\frac{4}{3}} \eta_{\min}^{*-\frac{8}{3}} \frac{1}{n} \Big( \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 \\ & - \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3 \Big) \Big( \sum_{k=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \Big) \\ \leq \quad & K^{-2} R^{\frac{4}{3}} \eta_{\min}^{*-\frac{8}{3}} \Big[ \frac{1}{n} \sum_{i=1}^{n} \Big( \sum_{k=1}^{K} 3K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^6 \\ & + 9K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^4 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^2 \\ & + 9K (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^4 \Big) \sum_{k'=1}^{K} (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^4 \Big]. \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}\Big( \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^4 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^4 \Big) \\ \leq C \|\boldsymbol{z}_k\|_2^2 \|\bar{\boldsymbol{\beta}}_k^*\|_2^4 \|\bar{\boldsymbol{\beta}}_{k'}\|_2^4. \end{aligned}$$

Together with Lemma 1 and (A.13), 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

$$\begin{aligned} B_1 \leq & K^{-2} R^{\frac{4}{3}} \eta_{\min}^{*-\frac{8}{3}} \Big[ 18C \Big( 1 + \frac{(\log n)^5}{\sqrt{n}} \Big) * \\ & K^2 \eta_{\max}^{*\frac{8}{3}} (1 + K^{\frac{1}{2}} \varepsilon_0)^4 \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2 \Big] \\ \leq \quad & R^4 18C \Big( 1 + \frac{(\log n)^5}{\sqrt{n}} \Big) (1 + K^{\frac{1}{2}} \varepsilon_0)^4 \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2, \end{aligned} \tag{A.62}$$

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

$$\begin{aligned} & \Big| \sum_{i=1}^{n} \epsilon_i^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 - \mathbb{E}\Big( \sum_{i=1}^{n} \epsilon_i^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \Big) \Big| \\ \leq \quad & C (\log n)^2 \|\boldsymbol{\epsilon}^2\|_2 \|\bar{\boldsymbol{\beta}}_k\|_2^4. \end{aligned}$$

Note that $\mathbb{E}((\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4) = 3 \|\bar{\boldsymbol{\beta}}_k\|_2^4$. It will reduce to

$$\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \leq \Big( \frac{3}{n} \sum_{i=1}^{n} \epsilon_i^2 + C \frac{(\log n)^2}{n} \|\boldsymbol{\epsilon}^2\|_2 \Big) \|\bar{\boldsymbol{\beta}}_k\|_2^4.$$

From Lemma 23, with probability at least $1 - 3/n$,

$$\Big|\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \Big| \leq C_0 \sigma^2, \;\; \frac{1}{n} \|\boldsymbol{\epsilon}^2\|_2 \leq C_0 \frac{\sigma^2}{\sqrt{n}}.$$

Combining the above two inequalities, we obtain

$$\Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4 \Big| \leq 6 C_0 \sigma^2 \|\bar{\boldsymbol{\beta}}_k\|_2^4, \tag{A.63}$$

with probability at least $1 - 7/n$. Plugging in the definition of $\phi$ and (A.13), $B_2$ is upper bounded by

$$B_2 \leq 6 C_0 \sigma^2 (1 + K^{\frac{1}{2}} \varepsilon_0)^4 \eta_{\min}^{*-\frac{4}{3}} R^{\frac{8}{3}} K^{-3}, \tag{A.64}$$

with probability at least $1 - 7K/n$.

**Summary.** Putting the bounds (A.62) and (A.64) together, we have similar upper bound for thresholded effect,

$$B \le C_2\mu^2 R^4 \sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2 + C_3\mu^2 \eta_{\min}^{*-\frac{4}{3}} R^{\frac{8}{3}} K^{-2}\frac{\sigma^2 s\log p}{n}, \tag{A.65}$$

with probability at least $1-(3K^2+7K)/n$. ∎

*3) Ensemble:* From the definition of $\gamma_k$, it's not hard to see actually the cross term $C$ is equal to zero. Combining the upper bound of gradient update effect (A.61) and thresholding effect (A.65) together, we obtain

$$\begin{aligned}&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \le \Big(1-64\mu K^{-2}R^{-\frac{8}{3}} \\ &+3\mu^2K^{-1}R^4[220+270K]^2\Big)\Big(\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2\Big) \\ &+2C_3\mu^2R^{\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n}.\end{aligned}$$

As long as the step size $\mu$ satisfies

$$0<\mu\le\frac{32\ R^{-20/3}}{3K[220+270K]^2},$$

we reach the conclusion

$$\begin{aligned}&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \\ &\le\Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^K\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \\ &\quad+2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n},\end{aligned} \tag{A.66}$$

with probability at least $1-4Ks/n$. ∎

### *M. Proof of Lemma 14*

Let us consider $k$-th component first. Without loss of generality, suppose $F\subset\{1,2,\ldots,Ks\}$. For $j=Ks+1,\ldots,p$,

$$\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)=\frac{2}{n}\sum_{i=1}^n\Big(\sum_{k=1}^K\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i\Big)\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2x_{ij}, \tag{A.67}$$

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 $1-\frac{1}{n^2p^2}$,

$$\begin{aligned}&\Big|\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)\Big|\le\frac{\sqrt{4\log(np)}}{n}* \\ &\sqrt{\sum_{i=1}^n(\sum_{k=1}^K\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i)^2(\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k))^2}=h(\boldsymbol{\beta}_k).\end{aligned}$$

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,

$$\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)\Big)=\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)_F\Big), \tag{A.68}$$

holds for $k\in[K]$, with probability at least $1-\frac{1}{n^2p}$. (A.68) 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. ∎

### *N. Proof of Lemma 15*

First, we consider symmetric case. According to the definition of $\{y_i\}_{i=1}^n$ from symmetric tensor estimation model (III.1), we separate the random noise $\epsilon_i$ by the following expansion,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n y_i^2 &= \frac{1}{n}\sum_{i=1}^n\Big[\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3+\epsilon_i\Big]^2 \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n(\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2}_{I_1} \\ &\quad+\underbrace{\frac{2}{n}\sum_{i=1}^n\epsilon_i\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3}_{I_2}+\underbrace{\frac{1}{n}\sum_{i=1}^n\epsilon_i^2}_{I_3}.\end{aligned} \tag{A.69}$$

**Bounding** $I_1$**.** We expand $i$-th component of $I_1$ as follows

$$\begin{aligned}&(\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2 \\ =&\sum_{k=1}^K\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6+2\sum_{k_i<k_j}\eta_{k_i}^*\eta_{k_j}^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3.\end{aligned} \tag{A.70}$$

As shown in Corollary A.2, the expectations of above two parts takes forms of

$$\begin{aligned}&\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^* \\ =\ &6(\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*)^3+9(\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*)\|\boldsymbol{\beta}_{k_i}^*\|_2^2\|\boldsymbol{\beta}_{k_j}^*\|_2^2 \\ &\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6=15\|\boldsymbol{\beta}_k^*\|_2^2.\end{aligned}$$

Recall that $\|\boldsymbol{\beta}_k^*\|_2=1$ for any $k\in[K]$ and Condition 3 implies for any $k_i\neq k_j$, $|\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*|\le\Gamma$, where $\Gamma$ 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

$$\Big|\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3\Big|\le 6\Gamma^3+9\Gamma,\ \text{for any } k_i\neq k_j \tag{A.71}$$

By using the concentration result in Lemma 1, we have with probability at least $1-1/n$

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6-\mathbb{E}(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6)\Big|\le C_1\frac{(\log n)^3}{\sqrt{n}}, \\ &\Big|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3-\mathbb{E}(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3)\Big| \\ &\le C_1\frac{(\log n)^3}{\sqrt{n}}.\end{aligned} \tag{A.72}$$

Putting (A.70),(A.71) and (A.72) together, this essentially provides an upper bound for $I_1$, namely

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2\\ \leq&\Big(15+6\Gamma^3+9\Gamma+2\ C_1\frac{(\log n)^3}{\sqrt{n}}\Big)(\sum_{k=1}^{K}\eta_k^*)^2, \quad \text{(A.73)}\end{aligned}$$

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

$$\begin{aligned}&\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3\geq C_2\frac{(\log n)^{\frac{3}{2}}}{n}\sqrt{n}\sigma\Big)\\ \leq\ &\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3\geq C_2\frac{\sigma(\log n)^{\frac{3}{2}}}{\sqrt{n}}\Big|\|\boldsymbol{\epsilon}\|_2\leq C_0\sigma\sqrt{n}\Big)\\ &+\mathbb{P}\Big(\|\boldsymbol{\epsilon}\|_2\geq C_0\sqrt{n}\sigma\Big)\\ \leq\ &\frac{1}{n}+\frac{3}{n}=\frac{4}{n}.\end{aligned}$$

This further implies that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3\epsilon_i\\ \leq\ &(\sum_{k=1}^{K}\eta_k^*)C_2\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}\sigma, \quad \text{(A.74)}\end{aligned}$$

with probability at least $1-4K/n$.

**Bounding** $I_3$**.** As shown in Lemma 23, the random noise $\epsilon_i$ with sub-exponential tail satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\leq C_3\sigma^2. \quad \text{(A.75)}$$

with probability at least $1-3/n$.

Overall, putting (A.73), (A.74) and (A.75) together, we have with probability at least $1-(K^2+4K+3)/n$,

$$\begin{aligned}&\frac{\frac{1}{n}\sum_{i=1}^n y_i^2}{(\sum_{k=1}^K\eta_k^*)^2}\leq 15+6\Gamma^3+9\Gamma+2\ C_1\frac{(\log n)^3}{\sqrt{n}}\\ &+\frac{2\ C_2\sigma}{(\sum_{k=1}^K\eta_k^*)}\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}+\frac{C_3\sigma^2}{(\sum_{k=1}^K\eta_k^*)^2}.\end{aligned}$$

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

$$\mathbb{E}(\boldsymbol{u}_i^\top\boldsymbol{\beta}_{1k}^*)^2(\boldsymbol{v}_i^\top\boldsymbol{\beta}_{2k}^*)^2(\boldsymbol{w}_i^\top\boldsymbol{\beta}_{3k}^*)^2=1.$$

■

### *O. Non-Symmetric Tensor Estimation*

*1) Conditions and Algorithm:* In this subsection, we provide several essential conditions for Theorem 5 and the detail algorithm for non-symmetric tensor estimation.

*Condition 6* (*Uniqueness of CP-Decomposition):* The CP-decomposition form (VI.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

$$\begin{aligned}\Gamma\ :=\ &\max_{k_i\neq k_j}\Big\{|\langle\boldsymbol{\beta}_{1k_i}^*,\boldsymbol{\beta}_{1k_j}^*\rangle|,|\langle\boldsymbol{\beta}_{2k_i}^*,\boldsymbol{\beta}_{2k_j}^*\rangle|,|\langle\boldsymbol{\beta}_{3k_i}^*,\boldsymbol{\beta}_{3k_j}^*\rangle|\Big\}\\ \leq\ &C\min\{K^{-\frac{3}{4}}R^{-1},s^{-\frac{1}{2}}\}.\end{aligned}$$

*Condition 9* (*Random Noise):* We assume the random noise $\{\epsilon_i\}_{i=1}^n$ follows a sub-exponential tail with parameter $\sigma$ satisfying $0<\sigma<C\sum_{k=1}^K\eta_k^*$.

*2) Proof of Theorem 5:* 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 (VI.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 $|\mathrm{supp}(\boldsymbol{\beta}_1^*)|=s_1$, $|\mathrm{supp}(\boldsymbol{\beta}_2^*)|=s_2$, $|\mathrm{supp}(\boldsymbol{\beta}_3^*)|=s_3$ and denote $s=\max\{s_1,s_2,s_3\}$. Define $F_j^{(t)}=\mathrm{supp}(\boldsymbol{\beta}_j^*)\cup\mathrm{supp}(\boldsymbol{\beta}_j^{(t)})$, $F^{(t)}=\cup_{j=1}^3F_j^{(t)}$ and the oracle estimator as

$$\widetilde{\boldsymbol{\beta}}_1^{(t+1)}=\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_1^{(t)})}\Big(\boldsymbol{\beta}_j^{(t)}-\frac{\mu}{\phi}\nabla_1\mathcal{L}(\boldsymbol{\beta}_1^{(t)},\boldsymbol{\beta}_2^{(t)},\boldsymbol{\beta}_3^{(t)})_{F^{(t)}}\Big),$$

where $h(\boldsymbol{\beta}_1^{(t)})$ has the form of

$$\begin{aligned}&\frac{\sqrt{4\log np}}{n}\Big(\sum_{i=1}^{n}\Big(\eta(\boldsymbol{u}_i^\top\boldsymbol{\beta}_1^{(t)})(\boldsymbol{v}_i^\top\boldsymbol{\beta}_2^{(t)})(\boldsymbol{w}_i^\top\boldsymbol{\beta}_3^{(t)})-y_i\Big)^2*\\ &\eta^{\frac{2}{3}}(\boldsymbol{v}_i^\top\boldsymbol{\beta}_2^{(t)})^2(\boldsymbol{w}_i^\top\boldsymbol{\beta}_3^{(t)})^2\Big)^{1/2}. \quad \text{(A.76)}\end{aligned}$$

The definitions of $\widetilde{\boldsymbol{\beta}}_2^{(t+1)}$ and $\widetilde{\boldsymbol{\beta}}_3^{(t+1)}$ are similar.

**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 $\mu$, rank $K$, stopping error $\epsilon = 10^{-4}$.
1: **Step 1:** Calculate the moment-based tensor $\mathcal{T}$ as (VI.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 (VI.5)
5: **Until** $\max\{\|\boldsymbol{B}_j^{(T+1)} - \boldsymbol{B}_j^{(T)}\|_F\} \le \epsilon$.
6: **Step 3:** Do column-wise normalization as

$$\widehat{\boldsymbol{B}}_1 = \Big(\frac{\boldsymbol{\beta}_{11}^{(T)}}{\|\boldsymbol{\beta}_{11}^{(T)}\|_2}, \dots, \frac{\boldsymbol{\beta}_{1K}^{(T)}}{\|\boldsymbol{\beta}_{1K}^{(T)}\|_2}\Big),$$
$$\widehat{\boldsymbol{B}}_2 = \Big(\frac{\boldsymbol{\beta}_{21}^{(T)}}{\|\boldsymbol{\beta}_{21}^{(T)}\|_2}, \dots, \frac{\boldsymbol{\beta}_{2K}^{(T)}}{\|\boldsymbol{\beta}_{2K}^{(T)}\|_2}\Big),$$
$$\widehat{\boldsymbol{B}}_3 = \Big(\frac{\boldsymbol{\beta}_{31}^{(T)}}{\|\boldsymbol{\beta}_{31}^{(T)}\|_2}, \dots, \frac{\boldsymbol{\beta}_{3K}^{(T)}}{\|\boldsymbol{\beta}_{3K}^{(T)}\|_2}\Big).$$

And update the weight by

$$\begin{aligned}\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}^{(0)} * (&\|\boldsymbol{\beta}_{11}^{(T)}\|_2\|\boldsymbol{\beta}_{21}^{(T)}\|_2\|\boldsymbol{\beta}_{31}^{(T)}\|_2, \\ &\dots, \|\boldsymbol{\beta}_{1K}^{(T)}\|_2\|\boldsymbol{\beta}_{3K}^{(T)}\|_2\|\boldsymbol{\beta}_{3K}^{(T)}\|_2)^\top.\end{aligned}$$

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 $\widehat{\mathscr{T}}$.

*Lemma 16:* Let $t \ge 0$ be an integer. Suppose Conditions 6-9 hold and $\{\boldsymbol{\beta}_j^{(t)}, \eta\}$ satisfies the following upper bound

$$\max_{j=1,2,3}\|\sqrt[3]{\eta}\boldsymbol{\beta}_j^{(t)} - \sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\|_2 \le \sqrt[3]{\eta^*}\varepsilon_0, \ |\eta - \eta^*| \le \varepsilon_0 \quad \text{(A.77)}$$

with probability at least $1 - CO(1/n)$. Assume the step size $\mu$ satisfies $0 < \mu < \mu_0$ for some small absolute constant $\mu_0$ and $s \le d \le Cs$. Then $\{\widetilde{\boldsymbol{\beta}}_j^{(t+1)}\}$ can be upper bounded as

$$\begin{aligned}&\max_{j=1,2,3}\Big\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_j^{(t+1)} - \sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\Big\|_2 \\ \le\ & (1 - \frac{\mu}{12})\max_{j=1,2,3}\Big\|\sqrt[3]{\eta}\boldsymbol{\beta}_j^{(t)} - \sqrt[3]{\eta^*}\boldsymbol{\beta}_j^*\Big\|_2 \\ &+\mu\frac{3\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{3s\log p}{n}},\end{aligned}$$

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

$$\begin{aligned}&\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3) \\ =\ & \sqrt[3]{\eta}\frac{2}{n}\sum_{i=1}^n\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\Big)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\boldsymbol{u}_i.\end{aligned}$$

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 $\boldsymbol{\gamma} \in \mathbb{R}^p$, $\mathrm{supp}(\boldsymbol{\gamma}) \subset F$ and $\|\boldsymbol{\gamma}\|_\infty \le 1$. 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,

$$\begin{aligned}&\Big\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\Big\|_2 \\ =&\underbrace{\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\Big\|_2}_{\text{gradient update effect}} + \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\Big|h(\bar{\boldsymbol{\beta}}_1)\Big|\sqrt{3s}}_{\text{thresholded effect}}.\end{aligned} \quad \text{(A.78)}$$

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 [55] given sufficient good initial. Following this proof strategy, we decompose the gradient update effect in (A.78) by three parts,

$$\begin{aligned}&\Big\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+ - \sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\Big\|_2 \\ \le\ & \underbrace{\Big\|\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_1^* - \mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)_F\Big\|_2}_{I_1} \\ &+ \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)_F - \nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\Big\|_2}_{I_2} \\ &+ \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F - \nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2, \bar{\boldsymbol{\beta}}_3)_F\Big\|_2}_{I_3} \\ &+ \underbrace{\mu\frac{\sqrt[3]{\eta}}{\phi}\Big|h(\bar{\boldsymbol{\beta}}_1)\Big|\sqrt{3s},}_{I_4}\end{aligned}$$

where $\nabla_1\widetilde{\mathcal{L}}$ is the noiseless gradient as we defined in (A.42). 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 L. And $\phi = \sum_{i=1}^n y_i^2$ approximates $\eta^{*2}$ up to constant due to Lemma 15.

*3) Bounding $I_1$:* In this section, let us denote

$$\begin{aligned}&\sqrt[3]{\eta}\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)/\phi = f(\bar{\boldsymbol{\beta}}_1) \\ &\sqrt[3]{\eta}\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2^*, \bar{\boldsymbol{\beta}}_3^*)/\phi = \nabla f(\bar{\boldsymbol{\beta}}_1),\end{aligned} \quad \text{(A.79)}$$

where $\mathrm{supp}(\nabla f(\bar{\boldsymbol{\beta}}_1)) = F$. 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 $F$,

$$\begin{aligned}&\nabla f(\bar{\boldsymbol\beta}_1^{(1)}) - \nabla f(\bar{\boldsymbol\beta}_1^{(2)})\\ =\;& \frac{(\sqrt[3]{\eta})^2}{\phi}\frac{2}{n}\sum_{i=1}^n \Big(\boldsymbol u_i^\top(\bar{\boldsymbol\beta}_1^{(1)}-\bar{\boldsymbol\beta}_1^{(2)})(\boldsymbol v_i^\top\bar{\boldsymbol\beta}_2^*)^2(\boldsymbol w_i^\top\bar{\boldsymbol\beta}_3^*)^2\Big)\boldsymbol u_i.\end{aligned}$$

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 (IV.7). Under Condition (5) with some constant adjustments, we obtain

$$\Big\|\nabla f(\bar{\boldsymbol\beta}_1^{(1)}) - \nabla f(\bar{\boldsymbol\beta}_1^{(2)})\Big\|_2 \le \frac{57}{16}\Big\|\bar{\boldsymbol\beta}_1^{(1)}-\bar{\boldsymbol\beta}_1^{(2)}\Big\|_2. \quad \text{(A.80)}$$

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 $\alpha$-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 [63], it shows that

$$\begin{aligned}&\lambda_{\min}\big(\nabla^2(f(\bar{\boldsymbol\beta}_1))\big)\\ \ge\;&\lambda_{\min}\big(\mathbb{E}(\nabla^2\ f(\bar{\boldsymbol\beta}_1))\big)-\lambda_{\max}\big(\nabla^2\ f(\bar{\boldsymbol\beta}_1)-\mathbb{E}(\nabla^2\ f(\bar{\boldsymbol\beta}_1))\big).\end{aligned} \quad \text{(A.81)}$$

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 $\lambda_{\min}\big(\mathbb{E}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\big)\ge 2$.

On the other hand,

$$\begin{aligned}&\lambda_{\max}\Big(\nabla^2 f(\bar{\boldsymbol\beta}_1)-\mathbb{E}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\Big)\\ =\;&\Big\|\nabla^2 f(\bar{\boldsymbol\beta}_1)-\mathbb{E}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\Big\|_2\\ \le\;&\boldsymbol a^\top\Big(\nabla^2 f(\bar{\boldsymbol\beta}_1)-\mathbb{E}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\Big)\boldsymbol b\\ =\;&\frac{2}{n}\sum_{i=1}^n(\boldsymbol v_i^\top\bar{\boldsymbol\beta}_2^*)^2(\boldsymbol w_i^\top\bar{\boldsymbol\beta}_3^*)^2(\boldsymbol u_i^\top\boldsymbol a)(\boldsymbol u_i^\top\boldsymbol b)\\ &-\mathbb{E}\Big(\sum_{i=1}^n(\boldsymbol v_i^\top\bar{\boldsymbol\beta}_2^*)^2(\boldsymbol w_i^\top\bar{\boldsymbol\beta}_3^*)^2(\boldsymbol u_i^\top\boldsymbol a)(\boldsymbol u_i^\top\boldsymbol b)\Big)\eta^{*-\frac{4}{3}}.\end{aligned}$$

where $\boldsymbol a,\boldsymbol b\in\mathbb{S}^{s-1}$. Equipped with Lemma 2, it yields that with probability at least $1-10/n^3$,

$$\lambda_{\max}\Big(\nabla^2 f(\bar{\boldsymbol\beta}_1)-\mathbb{E}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\Big)\le 2\delta_{n,s,p}.$$

Together with the lower bound of $\lambda_{\min}(\mathbb{E}(\nabla^2\ f(\bar{\boldsymbol\beta}_1)))$, we have

$$\lambda_{\min}(\nabla^2 f(\bar{\boldsymbol\beta}_1))\ge 2-2\delta_{n,p,s},$$

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 [55], it shows that

$$\begin{aligned}&\big(\nabla f(\bar{\boldsymbol\beta}_1)-\nabla f(\bar{\boldsymbol\beta}_1^*)^\top\big)\big(\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}^*\big)\\ \ge\;&\frac{\alpha L}{\alpha+L}\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*\Big\|_2^2+\frac{1}{\alpha+L}\Big\|\nabla f(\bar{\boldsymbol\beta}_1)-\nabla f(\bar{\boldsymbol\beta}_1^*)\Big\|_2^2.\end{aligned}$$

Since the gradient vanishes at the optimal point, the above inequality times $2\mu$ simplifies to

$$\begin{aligned}&-2\mu\nabla f(\bar{\boldsymbol\beta}_1)^\top(\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*)\\ &\le-\frac{2\mu\alpha L}{\alpha+L}\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*\Big\|_2^2-\frac{2\mu}{\alpha+L}\Big\|\nabla f(\bar{\boldsymbol\beta}_1)\Big\|_2^2.\end{aligned} \quad \text{(A.82)}$$

Now it's sufficient to bound $\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*-\mu\nabla f(\bar{\boldsymbol\beta}_1)\|_2$ as follows

$$\begin{aligned}&\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*-\mu\nabla f(\bar{\boldsymbol\beta}_1)\Big\|_2^2\\ =\;&\Big\|\bar{\boldsymbol\beta}_1^t-\bar{\boldsymbol\beta}_1^*\Big\|_2^2+\mu^2\Big\|\nabla f(\bar{\boldsymbol\beta}_1)\Big\|_2^2-2\mu\nabla f(\bar{\boldsymbol\beta}_1)^\top(\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}^*)\\ \le\;&\Big(1-2\mu\frac{\alpha L}{\alpha+L}\Big)\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*\Big\|_2^2\\ &+\mu(\mu-\frac{2}{\alpha+L})\Big\|\nabla f(\bar{\boldsymbol\beta}_1)\Big\|_2^2.\end{aligned}$$

where $L,\alpha$ 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,

$$\begin{aligned}&\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*-\mu\frac{\sqrt[3]{\eta}}{\phi}\nabla_1\widetilde{\mathcal{L}}\big(\bar{\boldsymbol\beta}_1,\bar{\boldsymbol\beta}_2^*,\bar{\boldsymbol\beta}_3^*\big)\Big\|_2\\ &\le\Big(1-3\mu\Big)\Big\|\bar{\boldsymbol\beta}_1-\bar{\boldsymbol\beta}_1^*\Big\|_2,\end{aligned} \quad \text{(A.83)}$$

with probability $1-20/n^3$. This ends the proof. ∎

*4) 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 $\varepsilon_0$ with probability at least $1-60/n^3$. Under Condition 5, it suffices to get

$$\begin{aligned}
&\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)-\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2^*,\bar{\boldsymbol{\beta}}_3^*)\Big\|_2\\
\leq&\frac{8}{3}\max_{j=1,2,3}\Big\|\bar{\boldsymbol{\beta}}_j-\bar{\boldsymbol{\beta}}_j^*\Big\|_2,
\end{aligned}\tag{A.84}$$

with probability at least $1-6/n$. ■

*5) 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

$$\begin{aligned}
&\mathbb{P}\Big(\max_{j\in[3s]}\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_i(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)u_{ij}\Big|\\
&\geq\ C(1+\varepsilon_0)^2\eta^{*-\frac{2}{3}}\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}\Big)\leq\frac{12s}{n}.
\end{aligned}$$

Therefore, we reach

$$\begin{aligned}
&\frac{\sqrt[3]{\eta}}{\phi}\Big\|\nabla_1\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)-\nabla_1\mathcal{L}(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2,\bar{\boldsymbol{\beta}}_3)\Big\|_2\\
\leq&2C\eta^{*-\frac{2}{3}}\sigma\sqrt{\frac{3s(\log n)^3}{n}},
\end{aligned}$$

with probability at least $1-12s/n$. ■

*6) Bounding $I_4$:* According to the definition of thresholding level $h(\boldsymbol{\beta}_1)$ in (A.76), we can bound the square as follows,

$$\begin{aligned}
&\frac{(\sqrt[3]{\eta})^2}{\phi^2}h^2(\bar{\boldsymbol{\beta}}_1)\\
=\ &\frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^n\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\\
&-(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)-\epsilon_i\Big)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2
\end{aligned}$$

Based on the basic inequality $(a+b)^2\leq 2(a^2+b^2)$, we have

$$\begin{aligned}
&\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\\
&-(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)-\epsilon_i\Big)^2\\
\leq\ &2\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\\
&-(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)\Big)^2+2\epsilon_i^2.
\end{aligned}$$

Denote $I_1$ and $I_2$ corresponding to optimization error and statistical error,

$$\begin{aligned}
I_1\ =\ &\frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^n\Big((\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)\\
&-(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)\Big)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
I_2\ =\ &\frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\sum_{i=1}^n\epsilon_i^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2.
\end{aligned}$$

Next, $I_1$ is decomposed by some high-order polynomials as follows

$$\begin{aligned}
I_1=&\frac{(\sqrt[3]{\eta})^4}{\phi^2}\frac{4\log np}{n^2}\Big(\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2\\
&(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
&+\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
&+\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
&+\sum_{i=1}^n(\boldsymbol{u}_i^\top\boldsymbol{z}_1)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2^*)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
&+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3^*)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\\
&+\sum_{i=1}^{n}(\boldsymbol{u}_i^\top\bar{\boldsymbol{\beta}}_1^*)^2(\boldsymbol{v}_i^\top\boldsymbol{z}_2)^2(\boldsymbol{w}_i^\top\boldsymbol{z}_3)^2(\boldsymbol{v}_i^\top\bar{\boldsymbol{\beta}}_2)^2(\boldsymbol{w}_i^\top\bar{\boldsymbol{\beta}}_3)^2\Big).
\end{aligned}
\tag{A.85}
$$

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 $1-7/n$,

$$
I_1\leq\frac{7\log p}{n}\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j-\bar{\boldsymbol{\beta}}_j^*\right\|_2^2. \tag{A.86}
$$

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,

$$
I_2\leq 4\sigma^2\eta^{*\frac{4}{3}}\frac{\log p}{n}. \tag{A.87}
$$

Putting (A.86) and (A.87) together, the thresholded effect can be bound by

$$
\begin{aligned}
\frac{\sqrt[3]{\eta}}{\phi}|h(\boldsymbol{\beta}_1)|\leq&\sqrt{\frac{7\log np}{n}}\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j-\bar{\boldsymbol{\beta}}_j^*\right\|_2\\
&+\frac{2\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{\log np}{n}},
\end{aligned}
\tag{A.88}
$$

with probability at least $1-8/n$, provided $n\gtrsim(\log n)^{10}$. ■

*7) Summary:* Putting the upper bounds (A.83), (A.84) and (A.88) together, we obtain that if step size $\mu$ satisfies $0<\mu<\mu_0$ for some small $\mu_0$,

$$
\begin{aligned}
\left\|\sqrt[3]{\eta}\widetilde{\boldsymbol{\beta}}_1^+-\sqrt[3]{\eta^*}\boldsymbol{\beta}_1^*\right\|_2\ \leq\ &(1-\frac{\mu}{12})\max_{j=1,2,3}\left\|\bar{\boldsymbol{\beta}}_j-\bar{\boldsymbol{\beta}}_j^*\right\|_2\\
&+\mu\frac{3\sigma}{(\sqrt[3]{\eta^*})^2}\sqrt{\frac{3s\log p}{n}},
\end{aligned}
$$

with probability at least $1-12s/n$. This finishes our proof. ■

### *P. Matrix Form Gradient and Stochastic Gradient Descent*

*1) Matrix Formulation of Gradient:* In this section, we provide detail derivations for (III.7) and (VI.5).

*Lemma A.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 (III.5) can be written in a matrix form as follows

$$
\begin{aligned}
\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B},\boldsymbol{\eta})\ =\ &\frac{6}{n}[((\boldsymbol{B}^\top\boldsymbol{X})^\top)^3\boldsymbol{\eta}-\boldsymbol{y}]^\top\\
&[(((\boldsymbol{B}^\top\boldsymbol{X})^\top)^2\odot\boldsymbol{\eta}^\top)^\top\odot\boldsymbol{X}]^\top.
\end{aligned}
$$

*Proof:* First let's have a look at the gradient for $k$-th component,

$$
\nabla\mathcal{L}_k(\boldsymbol{\beta}_k)=\frac{6}{n}(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i)\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)\boldsymbol{x}_i\in\mathbb{R}^{p\times 1},
$$

for $k=1,\ldots,K$. Correspondingly, each part can be written as a matrix form,

$$
\begin{aligned}
&((\underbrace{\boldsymbol{B}^\top\boldsymbol{X}}_{K\times n})^\top)^3\boldsymbol{\eta}-\boldsymbol{y}\in\mathbb{R}^{n\times 1}\\
&(((\boldsymbol{B}^\top\boldsymbol{X})^\top)^2\odot\boldsymbol{\eta}^\top)^\top\odot\boldsymbol{X}\in\mathbb{R}^{pK\times n}.
\end{aligned}
$$

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 (VI.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,

$$
\begin{aligned}
&\nabla_{\boldsymbol{B}_1}\mathcal{L}(\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{B}_3,\boldsymbol{\eta})=\boldsymbol{D}^\top(\boldsymbol{C}_1^\top\odot\boldsymbol{U})^\top\\
=&(\nabla_1\mathcal{L}(\boldsymbol{\beta}_1)^\top,\ldots,\nabla_K\mathcal{L}(\boldsymbol{\beta}_K)^\top).
\end{aligned}
$$

■

*2) Stochastic Gradient Descent:* Stochastic thresholded gradient descent is a stochastic approximation of the gradient descent optimization method. Note that the empirical risk function (III.5) that can be written as a sum of differentiable functions. Followed by (III.7), the gradient of (III.5) evaluated at $i$-th sketching $\{y_i,\boldsymbol{x}_i\}$ can be written as

$$
\begin{aligned}
\nabla_{\boldsymbol{B}}\mathcal{L}_i(\boldsymbol{B},\boldsymbol{\eta})\ =\ &[((\boldsymbol{B}^\top\boldsymbol{x}_i)^\top)^3\boldsymbol{\eta}-y_i]*\\
&[(((\boldsymbol{B}^\top\boldsymbol{x}_i)^\top)^2\odot\boldsymbol{\eta}^\top)^\top\odot\boldsymbol{x}_i]^\top\in\mathbb{R}^{1\times pK},
\end{aligned}
$$

Thus, the overall gradient $\nabla_{\boldsymbol{B}}\mathcal{L}_i(\boldsymbol{B},\boldsymbol{\eta})$ defined in (III.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

$$\begin{aligned}&\operatorname{vec}(\boldsymbol{B}^{(t+1)})\\ =\ &\varphi_{\frac{\mu_{SGD}}{\phi}\boldsymbol{h}(\boldsymbol{B}^{(t)})}\Big(\operatorname{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu_{SGD}}{\phi}\nabla_{\boldsymbol{B}}\mathcal{L}_i(\boldsymbol{B}^{(t)})\Big).\end{aligned}$$

### Q. 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,

$$\begin{aligned}&\mathbb{E}\Big((\boldsymbol{a}^\top\boldsymbol{x})(\boldsymbol{b}^\top\boldsymbol{x})(\boldsymbol{c}^\top\boldsymbol{x})\boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}\Big)\\ =&\Big(\boldsymbol{a}\circ\boldsymbol{b}\circ\boldsymbol{c}+\boldsymbol{a}\circ\boldsymbol{c}\circ\boldsymbol{b}+\boldsymbol{b}\circ\boldsymbol{a}\circ\\ &+\boldsymbol{b}\circ\boldsymbol{c}\circ\boldsymbol{a}+\boldsymbol{c}\circ\boldsymbol{b}\circ\boldsymbol{a}+\boldsymbol{c}\circ\boldsymbol{a}\circ\boldsymbol{b}\Big)\\ &+3\sum_{m=1}^p\Big(\boldsymbol{a}\circ\boldsymbol{e}_m\circ\boldsymbol{e}_m(\boldsymbol{b}^\top\boldsymbol{c})\\ &+\boldsymbol{e}_m\circ\boldsymbol{b}\circ\boldsymbol{e}_m(\boldsymbol{a}^\top\boldsymbol{c})+\boldsymbol{e}_m\circ\boldsymbol{e}_m\circ\boldsymbol{c}(\boldsymbol{a}^\top\boldsymbol{b})\Big),\end{aligned}\quad\text{(A.89)}$$

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 (A.89) 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

$$\mathbb{E}\Big((\boldsymbol{a}^\top\boldsymbol{u})(\boldsymbol{b}^\top\boldsymbol{v})(\boldsymbol{c}^\top\boldsymbol{w})\boldsymbol{u}\circ\boldsymbol{v}\circ\boldsymbol{w}\Big)=\boldsymbol{a}\circ\boldsymbol{b}\circ\boldsymbol{c}.\quad\text{(A.90)}$$

*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 (A.89), 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 A.2:* 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 $\psi_2$-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,

$$\begin{aligned}&\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n\Big[\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i-\mathbb{E}(\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i)\Big]\Big\|_s\\ &\qquad\qquad\geq CK_xK_yK_z\delta_{n,p,s}\Big)\leq p^{-1}.\\ &\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n\Big[\boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i-\mathbb{E}(\boldsymbol{x}_i\circ\boldsymbol{x}_i\circ\boldsymbol{x}_i)\Big]\Big\|_s\\ &\qquad\qquad\geq CK_x^3\delta_{n,p,s}\Big)\leq p^{-1}.\end{aligned}$$

Here, $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (II.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 $\epsilon$-net. Since we will bound a sparse tensor spectral norm, our strategy is to discrete the sparse set and construct the $\epsilon$-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 (II.3), we have

$$\begin{aligned}A&=\Big\|\frac{1}{n}\sum_{i=1}^n\Big[\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i-\mathbb{E}(\boldsymbol{x}_i\circ\boldsymbol{y}_i\circ\boldsymbol{z}_i)\Big]\Big\|_s\\ &=\sup_{\boldsymbol{\chi}_1,\boldsymbol{\chi}_2,\boldsymbol{\chi}_3\in\mathcal{B}_0}\Big|\frac{1}{n}\sum_{i=1}^n\Big[\langle\boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle\boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle\boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle\\ &\quad-\mathbb{E}(\langle\boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle\boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle\boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle)\Big]\Big|.\end{aligned}$$

Instead of constructing the $\epsilon$-net on $\mathcal{B}_0$, we will construct an $\epsilon$-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 [64], the cardinality of $\mathcal{N}_{0,s}$ is bounded by $5^s$. By Lemma 21, we obtain

$$\begin{aligned}&\sup_{\chi_1,\chi_2,\chi_3\in\mathcal{B}_{0,s}}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\langle \boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle \boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle \boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle\\ &-\mathbb{E}(\langle \boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle \boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle \boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle)\Big]\Big|\\ \leq & 2^3\sup_{\chi_1,\chi_2,\chi_3\in\mathcal{N}_{\mathcal{B}_{0,s}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\langle \boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle \boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle \boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle\\ &-\mathbb{E}(\langle \boldsymbol{x}_i,\boldsymbol{\chi}_1\rangle\langle \boldsymbol{y}_i,\boldsymbol{\chi}_2\rangle\langle \boldsymbol{z}_i,\boldsymbol{\chi}_3\rangle)\Big]\Big|. \qquad \text{(A.91)}\end{aligned}$$

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 $\psi_2$-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 (A.91) can be bounded by

$$\text{RHS} \geq 8\ K_xK_yK_zC\Big(\sqrt{\frac{\log\delta^{-1}}{n}}+\sqrt{\frac{(\log\delta^{-1})^3}{n^2}}\Big),$$

with probability smaller than $(5^s)^3\delta$ for any $0<\delta<1$.

Lastly, taking the union bound over all possible subsets $\mathcal{B}_{0,s}$ yields that

$$\begin{aligned}&\mathbb{P}\Big(A\geq 8\ K_xK_yK_zC\Big(\sqrt{\frac{\log\delta^{-1}}{n}}+\sqrt{\frac{(\log\delta^{-1})^3}{n^2}}\Big)\Big)\\ \leq\ &(\frac{ep}{s})^s(5^s)^3\delta=(\frac{125ep}{s})^s\delta.\end{aligned}$$

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 [65])):* Let $\mathbb{N}$ be an $\epsilon$-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 $\epsilon$-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

$$\begin{aligned}&\mathbb{P}\Big(\big|X_1X_2\big|>t\Big)=\mathbb{P}\Big(\big|X_2\big|>\frac{t}{\big|X_1\big|}\Big)\\ \leq\ &\mathbb{P}\Big(\big|X_2\big|>\frac{t}{\big|K_1\big|}\Big)\leq\exp\Big(1-t^2/K_1^2K_2^2\Big),\end{aligned}$$

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 [48])):* Suppose $\epsilon_1,\ldots,\epsilon_n$ are independent centered sub-exponential random variables with

$$\sigma:=\max_{1\leq i\leq n}\|\epsilon_i\|_{\psi_1}.$$

Then with probability at least $1-3/n$, we have

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\Big|\leq C_0\sigma\sqrt{\frac{\log n}{n}},\ \big\|\boldsymbol{\epsilon}\big\|_\infty\leq C_0\sigma\log n,\\ &\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\Big|\leq C_0\sigma^2,\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^4\Big|\leq C_0\sigma^4,\end{aligned}$$

for some constant $C_0$.

*Lemma 24 (Tail Probability for the Sum of Weibull Distributions (Lemma 3.6 in [34])):* Let $\alpha\in[1,2]$ and $Y_1,\ldots,Y_n$ be independent symmetric random variables satisfying $\mathbb{P}(|Y_i|\geq t)=\exp(-t^\alpha)$. Then for every vector $\boldsymbol{a}=(a_1,\ldots,a_n)\in\mathbb{R}^n$ and every $t\geq 0$,

$$\mathbb{P}\Big(|\sum_{i=1}^{n}a_iY_i|\geq t\Big)\leq 2\exp\Big(-c\min\Big(\frac{t^2}{\|\boldsymbol{a}\|_2^2},\frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big)$$

*Proof:* It is a combination of Corollaries 2.9 and 2.10 in [58].

*Lemma 25 (Moments for the Sum of Weibull Distributions (Corollary 1.2 in [66])):* Let $X_1,X_2,\ldots,X_n$ be a sequence of independent symmetric random variables satisfying $\mathbb{P}(|Y_i|\geq t)=\exp(-t^\alpha)$, where $0<\alpha<1$. Then, for $p\geq 2$ and some constant $C(\alpha)$ which depends only on $\alpha$,

$$\Big\|\sum_{i=1}^{n}a_iX_i\Big\|_p\leq C(\alpha)(\sqrt{p}\|\boldsymbol{a}\|_2+p^{1/\alpha}\|\boldsymbol{a}\|_\infty).$$

*Lemma 26 (Stein's Lemma [56]):* 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 [67])):* 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

$$\Big\|\sum_{i=1}^{n}\varepsilon_ia_i\Big\|_q\leq\Big(\frac{q-1}{p-1}\Big)^{1/2}\Big\|\sum_{i=1}^{n}\varepsilon_ia_i\Big\|_p.$$

*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 $\mathcal{S}_{k_1k_2} \subset [1, 2, \dots, p]$ as an index set such that for any $i, j \in \mathcal{S}_{k_1k_2}$, we have $x_{k_1 i} \neq 0$ and $x_{k_2 j} \neq 0$. From the definition of $\mathcal{S}_{k_1k_2}$, we know that $|\mathcal{S}_{k_1k_2}| \leq s$ and $\boldsymbol{x}_{k_1}^\top \boldsymbol{x}_{k_2} = \sum_{j=1}^p x_{k_1 j}x_{k_2 j} = \sum_{j\in\mathcal{S}_{k_1k_2}} x_{k_1 j}x_{k_2 j}$. We apply standard Hoeffding's concentration inequality,

$$\begin{aligned}&\mathbb{P}\Big(|\langle \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}\rangle| \geq t\Big)\\ = \ &\mathbb{P}\Big(|\sum_{j\in\mathcal{S}_{k_1k_2}} x_{k_1 j}x_{k_2 j}| \geq t\Big) \leq e\exp\Big(-\frac{ct^2}{s}\Big).\end{aligned}$$

Letting $ct^2/s = \log(1/\delta)$, we reach the conclusion.

## Acknowledgment

While completing this work, Guang Cheng was a member of the Institute for Advanced Study, Princeton, and a Visiting Fellow of SAMSI for the Deep Learning Program in the Fall of 2019; he would like to thank both Institutes for their hospitality.

## References

[1] P. M. Kroonenberg, *Applied Multiway Data Analysis* (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2008.
[2] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.
[3] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Amer. Stat. Assoc.*, vol. 108, no. 502, pp. 540–552, Jun. 2013.
[4] X. Li, D. Xu, H. Zhou, and L. Li, "Tucker tensor regression and neuroimaging analysis," *Statist. Biosci.*, vol. 10, no. 3, pp. 520–545, Dec. 2018.
[5] W. W. Sun and L. Li, "Store: Sparse tensor response regression and neuroimaging analysis," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4908–4944, 2017.
[6] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdiscip. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 6, pp. 355–380, Nov. 2013.
[7] S. Friedland, Q. Li, and D. Schonfeld, "Compressive sensing of sparse tensors," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4438–4447, Oct. 2014.
[8] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.
[9] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Found. Comput. Math.*, vol. 16, no. 4, pp. 1031–1068, Aug. 2016.
[10] M. Yuan and C.-H. Zhang, "Incoherent tensor norms and their applications in higher order tensor completion," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6753–6766, Oct. 2017.
[11] A. Zhang, "Cross: Efficient low-rank tensor completion," *Ann. Statist.*, vol. 47, no. 2, pp. 936–964, Apr. 2019.
[12] A. Montanari and N. Sun, "Spectral algorithms for tensor completion," *Commun. Pure Appl. Math.*, vol. 71, no. 11, pp. 2381–2425, Nov. 2018.
[13] N. Ghadermarzy, Y. Plan, and Ö. Yilmaz, "Near-optimal sample complexity for convex tensor completion," *Inf. Inference, J. IMA*, vol. 8, no. 3, pp. 577–619, Sep. 2019.
[14] Z. Zhang and S. Aeron, "Exact tensor completion using t-SVD," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1511–1526, Mar. 2017.
[15] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do, "Efficient tensor completion for color image and video recovery: Low-rank tensor train," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2466–2479, May 2017.
[16] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multiresponse tensor regression," *Ann. Statist.*, vol. 47, no. 3, pp. 1554–1584, Jun. 2019.
[17] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 172–208, 2019.
[18] L. Li and X. Zhang, "Parsimonious tensor response regression," *J. Amer. Stat. Assoc.*, vol. 112, no. 519, pp. 1131–1146, Jul. 2017.
[19] A. Zhang, Y. Luo, G. Raskutti, and M. Yuan, "ISLET: Fast and optimal low-rank tensor regression via importance sketching," 2019, *arXiv:1911.03804*. [Online]. Available: http://arxiv.org/abs/1911.03804
[20] B. Romera-Paredes, M. H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, vol. 28, 2013, pp. III-1444–III-1452.
[21] J. Bien, J. Taylor, and R. Tibshirani, "A lasso for hierarchical interactions," *Ann. Statist.*, vol. 41, no. 3, pp. 1111–1141, Jun. 2013.
[22] N. Hao and H. H. Zhang, "Interaction screening for ultrahigh-dimensional data," *J. Amer. Stat. Assoc.*, vol. 109, no. 507, pp. 1285–1301, Jul. 2014.
[23] Y. Fan, Y. Kong, D. Li, and J. Lv, "Interaction pursuit with feature screening and selection," 2016, *arXiv:1605.08933*. [Online]. Available: http://arxiv.org/abs/1605.08933
[24] S. Basu, K. Kumbier, J. B. Brown, and B. Yu, "Iterative random forests to discover predictive and stable high-order interactions," *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 8, pp. 1943–1948, Feb. 2018.
[25] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 517–520.
[26] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2003, p. II-93.
[27] W. W. Sun, J. Lu, H. Liu, and G. Cheng, "Provable sparse tensor decomposition," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 79, no. 3, pp. 899–916, Jun. 2017.
[28] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra Appl.*, vol. 523, pp. 220–262, Jun. 2017.
[29] X. Li, J. Haupt, and D. Woodruff, "Near optimal sketching of low-rank tensor regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3466–3476.
[30] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Ann. Statist.*, vol. 42, no. 6, pp. 2164–2201, Dec. 2014.
[31] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with non-convexity: Statistical and algorithmic theory for local optima," *J. Mach. Learn. Res.*, vol. 16, pp. 559–616, Mar. 2015.
[32] T. T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *Ann. Statist.*, vol. 43, no. 1, pp. 102–138, Feb. 2015.
[33] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 4034–4059, Jul. 2015.
[34] 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 Approx.*, vol. 34, no. 1, pp. 61–88, Aug. 2011.
[35] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, 2009.
[36] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.
[37] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, Oct. 2011.
[38] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2897–2905.
[39] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338, Nov. 2018.
[40] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, E. P. Xing and T. Jebara, Eds. Bejing, China, Jun. 2014, pp. 73–81.
[41] S. Friedland and L.-H. Lim, "Nuclear norm of higher-order tensors," *Math. Comput.*, vol. 87, no. 311, pp. 1255–1281, 2018.

[42] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, p. 45, 2013.

[43] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, Aug. 2014.

[44] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[45] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, Apr. 2011.

[46] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 27th Conf. Learn. Theory*, vol. 35, M. F. Balcan, V. Feldman, and C. Szepesvári, Eds., Barcelona, Spain, Jun. 2014, pp. 779–806.

[47] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing" *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3537–3580, 2016.

[48] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow," *Ann. Statist.*, vol. 44, no. 5, pp. 2221–2251, Oct. 2016.

[49] M. Janzamin, H. Sedghi, and A. Anandkumar, "Score function features for discriminative learning: Matrix and tensor framework," 2014, *arXiv:1412.2863*. [Online]. Available: http://arxiv.org/abs/1412.2863

[50] A. Anandkumar, R. Ge, and M. Janzamin, "Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates," 2014, *arXiv:1402.5180*. [Online]. Available: http://arxiv.org/abs/1402.5180

[51] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank symmetric tensor completion from noisy data," 2019, *arXiv:1911.04436*. [Online]. Available: http://arxiv.org/abs/1911.04436

[52] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[53] 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, Mar. 2016.

[54] N. D. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 757–760, Nov. 2012.

[55] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, nos. 3–4, pp. 231–357, 2015.

[56] C. Stein, P. Diaconis, S. Holmes, and G. Reinert, "Use of exchangeable pairs in the analysis of simulations," in *Stein's Method*. Institute of Mathematical Statistics, 2004, pp. 1–25.

[57] P. Hitczenko, S. J. 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.

[58] M. Talagrand, "The supremum of some canonical processes," *Amer. J. Math.*, vol. 116, no. 2, pp. 283–325, Apr. 1994.

[59] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing*. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 210–268.

[60] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[61] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2013.

[62] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, M. F. Balcan and K. Q. Weinberger, Eds. New York, NY, USA, Jun. 2016, pp. 964–973.

[63] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY, USA: Cambridge Univ. Press, 1988.

[64] M. Ledoux, *The Concentration of Measure Phenomenon*, vol. 89. Providence, RI, USA: American Mathematical Society, 2005.

[65] N. H. Nguyen, P. Drineas, and T. D. Tran, "Tensor sparsification via a bound on the spectral norm of random tensors," *Inf. Inference, J. IMA*, vol. 4, no. 3, pp. 195–229, 2015.

[66] R. Bogucki, "Suprema of canonical weibull processes," *Statist. Probab. Lett.*, vol. 107, pp. 253–263, Dec. 2015.

[67] V. De la Pena and E. Giné, *Decoupling: From Dependence to Independence*. Springer, 2012.

**Botao Hao** received the B.S. degree from the School of Mathematics, Nankai University, China, in 2014, and the Ph.D. degree from the Department of Statistics, Purdue University, USA, in 2019. He is currently a Post-Doctoral Researcher with the Department of Electrical Engineering, Princeton University, USA.

**Anru Zhang** received the B.S. degree from Peking University, Beijing, China, in 2010, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, PA, USA, in 2015. He is currently an Assistant Professor of statistics with the University of Wisconsin–Madison, Madison, WI, USA. His current research interests include high-dimensional statistical inference, tensor data analysis, statistical learning theory, dimension reduction, and convex/nonconvex optimization.

**Guang Cheng** received the B.A. degree in economics from Tsinghua University, China, in 2002, and the Ph.D. degree from the University of Wisconsin–Madison in 2006. He then joined the Department of Statistics, Duke University, as a Visiting Assistant Professor and a Post-Doctoral Fellow in SAMSI. He is currently a Professor of statistics with Purdue University, directing the Big Data Theory Research Group, whose main goal is to develop computationally efficient inferential tools for big data with statistical guarantees.