Supplement to "Sparse and Low-rank Tensor Estimation via Cubic Sketchings"

This supplementary contains five parts: (1) Section A contains high-order interaction effect model using our cubic sketching framework; (2) Section B includes detailed proofs for empirical moment estimator and concentration results; (3) Section C provides additional proofs for the main theoretical results of this paper; (2) Section D contains detailed proofs for the theoretical developments in the main theorems; (4) Section E discusses the matrix form of gradient function and stochastic gradient descent; (5) Section F provides several technical lemmas and their proofs.

A Application to High-Order Interaction Effect Models

In this section, we estimate high-order interaction effect models in the cubic sketching framework (see Figure 3). Specifically, we consider the following three-way interaction model

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

Here $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\eta}$ are coefficients for main effect, pairwise interaction, and triple-wise interaction, respectively. Importantly, (S.1) can be reformulated as the following tensor form (also see the left panel in Figure 3)

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

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:

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

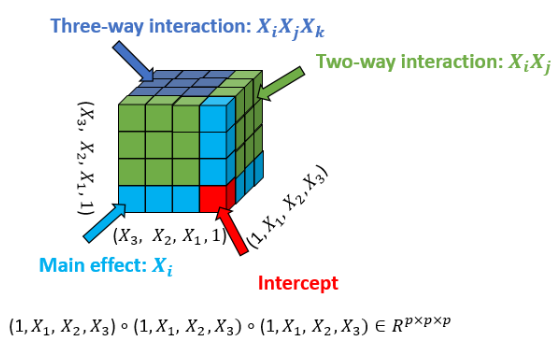

Figure 3: Illustration for interaction reformulation.

We next argue that it is reasonable to assume $\mathcal{B}$ is low rank and sparse in the tensor formulation of high-order interaction models. First, in modern biomedical research such as Hung et al. (2016), only a small portion of coefficients contribute to the response, leading to a highly sparse $\mathcal{B}$. Further, Sidiropoulos and Kyrillidis (2012) suggested that for the low-enough rank it is suitable to model sparse tensors as arising from sparse loadings, saying CP-decomposition. Moreover, this low-rank-and-sparse assumption (or approximation) seems necessary when the sample size is limited. Specifically, we assume $\mathcal{B}$ is of CP rank-K with s-sparse factors, where $K, s \ll p$. It is easy to see that the number of parameters in (S.4) is $K(p+1)$, which is significantly smaller than $(p+1)^3$, the total number of parameters in the original three-way interaction effect model (S.1). In this case, (S.2) can be written as

$$\begin{aligned} &y_l = \Big\langle \sum_{k=1}^{K} \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \boldsymbol{x}_l \circ \boldsymbol{x}_l \circ \boldsymbol{x}_l \Big\rangle + \epsilon_l, \quad l = 1, \dots, n, \\ &\text{where} \quad \|\boldsymbol{\beta}_k\|_2 = 1, \quad \|\boldsymbol{\beta}_k\|_0 \le s, \quad k \in [K]. \end{aligned} \tag{S.4}$$

By assuming $\boldsymbol{z}_l \overset{iid}{\sim} N_p(0, \boldsymbol{I}_p)$, the high-order interaction effect model (S.2) reduces to the symmetric tensor estimation model (3.1) with the only difference that the first coordinate of $\boldsymbol{x}_l$, i.e., the intercept, is always 1. To accommodate this slight difference, we only need to adjust the initial unbiased estimate in the above two-step procedure. We first obtain $\mathcal{T}_s$ by replacing $\boldsymbol{x}_i$ therein by $\boldsymbol{x}_l$, where $\boldsymbol{x}_l$ corresponds the l-th observation

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

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

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

Lemma 4 verifies that $\mathcal{T}_{s'}$ is an unbiased estimator for $\mathcal{B}$.

Theoretical results in Section 4 imply the following upper and lower bound results in this particular example.

Corollary 1 . Suppose that $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ 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

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

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

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

then the following lower bound holds,

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

B Main Proofs

In this section, we provide detailed proofs for empirical moment estimator and concentration results in Sections S.I and S.II.

S.I Moment Calculation

We first introduce three lemmas to show that the empirical moment based tensors are all unbiased estimators for the target low-rank tensor in the corresponding scenarios. Detail proofs of three lemmas are postponed to Sections S.I.1, S.I.2 and S.I.3 in the supplementary materials.

Lemma 2 (Unbiasedness of moment estimator under non-symmetric sketchings). Consider a non-symmetric tensor estimation model as follows

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

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

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

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 Janzamin et al. (2014). To fix the idea, we present only third order result for simplicity. The extension to higher-order is straightforward.

Theorem 5 (Third-order Stein's Identity, (Janzamin et al., 2014)). Let $\boldsymbol{x} \in \mathbb{R}^p$ be a random vector with joint density function $p(\boldsymbol{x})$. Define the third order score function $\mathcal{S}_3(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}^{p \times p \times p}$ as $\mathcal{S}_3(\boldsymbol{x}) = -\nabla^3 p(\boldsymbol{x}) / p(\boldsymbol{x})$. Then for continuously differentiable function $G(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}$, we have

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

In general, the order-m high-order score function is defined as

$$\mathcal{S}_m(\boldsymbol{x}) = (-1)^m \frac{\nabla^m p(\boldsymbol{x})}{p(\boldsymbol{x})}.$$

Interestingly, the high-order score function has a recursive differential representation

$$\mathcal{S}_m(\boldsymbol{x}) := -\mathcal{S}_{m-1}(\boldsymbol{x}) \circ \nabla \log p(\boldsymbol{x}) - \nabla \mathcal{S}_{m-1}(\boldsymbol{x}), \tag{S.4}$$

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 24 (Stein's lemma (Stein et al., 2004)). The proof of Theorem 5 relies on iteratively applying the recursion representation of score function (S.4) and the first-order Stein's lemma (Lemma 24). We provide the detailed proof in Section S.IV for the sake of completeness.

In particular, if $\boldsymbol{x}$ follows a standard Gaussian vector, each order score function can be calculated based on (S.4) as follows,

$$\begin{aligned} &\mathcal{S}_1(\boldsymbol{x}) = \boldsymbol{x}, \mathcal{S}_2(\boldsymbol{x}) = \boldsymbol{x}\circ\boldsymbol{x} - I_{d\times d}, \\ &\mathcal{S}_3(\boldsymbol{x}) = \boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x} - \sum_{j=1}^{p}\Big(\boldsymbol{x}\circ\boldsymbol{e}_j\circ\boldsymbol{e}_j + \boldsymbol{e}_j\circ\boldsymbol{x}\circ\boldsymbol{e}_j + \boldsymbol{e}_j\circ\boldsymbol{e}_j\circ\boldsymbol{x}\Big). \end{aligned} \tag{S.5}$$

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

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

which is exactly $\mathscr{T}^*$. Connecting this fact with (S.3), we are able to construct the unbiased estimator in the following lemma through high-order Stein's identity.

Lemma 3 (Unbiasedness of moment estimator under symmetric sketchings). Consider the symmetric tensor estimation model (3.1) & (4.8). 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 (S.5). By using the conclusion in Theorem 5 and the fact (S.6), we obtain

$$\mathbb{E}(\mathcal{T}_s) = \mathbb{E}\Big(\frac{1}{6n}\sum_{i=1}^{n}y_i\mathcal{S}_3(\boldsymbol{x})\Big) = \sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*,$$

since ϵ_i is independent of $\boldsymbol{x}_i$. This ends the proof. ■

Although the interaction effect model (S.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 4 (Unbiasedness of moment estimator in interaction model). For interaction effect model (S.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.,

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

S.II Proof of Lemma 1: Concentration Inequalities

We aim to prove Lemma 1 in this subsection. This lemma provides key concentration inequalities of the theoretical analysis for the main result. Before going into technical details, we introduce a quasi-norm called ψ_α-norm.

Definition 1 (ψ_α-norm (Adamczak et al., 2011)). The ψ_α-norm of any random variable X and $\alpha > 0$ is defined as

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

Particularly, a random variable who has a bounded ψ_2-norm or bounded ψ_1-norm is called sub-Gaussian or sub-exponential random variable, respectively. Next lemma provides an upper bound for the p-th moment of sum of random variables with bounded ψ_α-norm.

Lemma 5 . 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 $p \ge 2$,

$$\begin{aligned} &\Big(\mathbb{E}\Big|\sum_{i=1}^n a_i X_i - \mathbb{E}(\sum_{i=1}^n a_i X_i)\Big|^p\Big)^{\frac{1}{p}} \\ \le\ & \begin{cases} C_1(\alpha) b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty\big), & \text{if } 0 < \alpha < 1; \\ C_2(\alpha) b\big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\big), & \text{if } \alpha \ge 1. \end{cases} \end{aligned} \tag{S.7}$$

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

If $0 < \alpha < 1$, (S.7) is a combination of Theorem 6.2 in Hitczenko et al. (1997) and the fact that the p-th moment of a Weibull variable with parameter α is of order $p^{1/\alpha}$. If $\alpha \ge 1$, (S.7) follows from a combination of Corollaries 2.9 and 2.10 in Talagrand (1994). Continuing with standard symmetrization arguments, we reach the conclusion for general random variables. When $\alpha = 1$ or 2, (S.7) coincides with standard moment bounds for a sum of sub-Gaussian and sub-exponential random variables in Vershynin (2012). The detailed proof of Lemma 5 is postponed to Section S.II.

When $0 < \alpha < 1$, by Chebyshev's inequality, one can obtain the following exponential tail bound for the sum of random variables with bounded ψ_α-norm. This lemma generalizes the Hoeffding-type concentration inequality for sub-Gaussian random variables (see, e.g. Proposition 5.10 in Vershynin (2012)), and Bernstein-type concentration inequality for sub-exponential random variables (see, e.g. Proposition 5.16 in Vershynin (2012)).

Lemma 6 . 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 α such that for any $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $0 < \delta < 1/e^2$,

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

with probability at least $1 - \delta$.

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

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

where the last inequality is from Lemma 5. 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 $0 < \delta < 1/e^2$,

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

holds with probability at least $1 - \delta$. This ends the proof. ■

The next lemma provides an upper bound for the product of random variables in ψ_α-norm.

Lemma 7 (ψ_α for product of random variables). Suppose $X_1, \ldots, X_m$ are m random variables (not necessarily independent) with ψ_α-norm bounded by $\|X_j\|_{\psi_\alpha} \leq K_j$. Then the $\psi_{\alpha/m}$-norm of $\prod_{j=1}^m X_j$ is bounded as

$$\Big\|\prod_{j=1}^m X_j\Big\|_{\psi_{\alpha/m}} \leq \prod_{j=1}^m K_j.$$

Proof. For any $\{x_j\}_{j=1}^m$ and $\alpha > 0$, by using the inequality of arithmetic and geometric means we have

$$\Big(|\prod_{j=1}^m \frac{x_j}{K_j}|\Big)^{\alpha/m} = \Big(\prod_{j=1}^m |\frac{x_j}{K_j}|^\alpha\Big)^{1/m} \leq \frac{1}{m}\sum_{j=1}^m |\frac{x_j}{K_j}|^\alpha.$$

Since exponential function is a monotone increasing function, it shows that

$$\begin{aligned} &\exp\Big(|\prod_{j=1}^m \frac{x_j}{K_j}|\Big)^{\alpha/m} \leq \exp\Big(\frac{1}{m}\sum_{j=1}^m |\frac{x_j}{K_j}|^\alpha\Big) \\ &= \Big(\prod_{j=1}^m \exp(|\frac{x_j}{K_j}|^\alpha)\Big)^{1/m} \leq \frac{1}{m}\sum_{j=1}^m \exp\Big(|\frac{x_j}{K_j}|^\alpha\Big). \end{aligned} \tag{S.8}$$

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

$$\mathbb{E}\Big(\exp(\frac{|X_j|}{K_j})^\alpha\Big) \leq 2. \tag{S.9}$$

Putting (S.8) and (S.9) 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 $\prod_{j=1}^m K_j$. ■

Lemma 8 (Concentration inequality for sum of sub-Gaussian products). Suppose $\boldsymbol{X}_i = (\boldsymbol{x}_{1i}^\top, \ldots, \boldsymbol{x}_{mi}^\top)^\top \in \mathbb{R}^{m\times p}$, $i \in [n]$ are n i.i.d random matrices. Here, $\boldsymbol{x}_{ij}$ is the j-th row of $\boldsymbol{X}_i$ and suppose it is an isotropic sub-Gaussian vector. Then for any vectors $\boldsymbol{a} = (a_1 \ldots, a_n) \in \mathbb{R}^n$, $\{\boldsymbol{\beta}_j\}_{j=1}^m \subseteq \mathbb{R}^p$, and $0 < \delta < 1$, we have

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

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

Note that in Lemma 8, entries in each matrix $\boldsymbol{X}_i$ are not necessarily independent even $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent matrices.

Proof of Lemma 8. Note that for any $j=1,2,\dots,m$, the ψ_2-norm of $\boldsymbol{X}_j^\top\boldsymbol{\beta}_j$ is bounded by $\|\boldsymbol{\beta}_j\|_2$ (Vershynin, 2012). According to Lemma 7, 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 6, we reach the conclusion. ■

Proof of Lemma 1. We first start from the non-symmetric version in (S.1) 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 18 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 $\boldsymbol{e}_1=(1,0,\dots,0)^\top$. Then, it is equivalent to prove

$$\begin{aligned}&\Big\|M_{\rm nsy}-\mathbb{E}(M_{\rm nsy})\Big\|_s=\Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}x_{2i}x_{3i}\boldsymbol{x}_{1i}\circ\boldsymbol{x}_{2i}\circ\boldsymbol{x}_{3i}-\boldsymbol{e}_1\circ\boldsymbol{e}_1\circ\boldsymbol{e}_1\Big\|_s\\ \leq\quad & C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}}+\sqrt{\frac{s\log(p/s)}{n}}\Big).\end{aligned}$$

Suppose $\boldsymbol{x}_1\sim\mathcal{N}(0,\boldsymbol{I}_{p_1}),\boldsymbol{x}_2\sim\mathcal{N}(0,\boldsymbol{I}_{p_2}),\boldsymbol{x}_3\sim\mathcal{N}(0,\boldsymbol{I}_{p_3})$ and $\{\boldsymbol{x}_{1i},\boldsymbol{x}_{2i},\boldsymbol{x}_{3i}\}_{i=1}^n$ are n independent samples of $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$. And define a bounded event $\mathcal{G}_n$ for the first coordinate and its corresponding population version,

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

where M is a large constant to be specified later. Decomposing $\|M_{\rm nsy}-\mathbb{E}(M_{\rm nsy})\|_s$ as

$$\begin{aligned}&\Big\|M_{\rm nsy}-\mathbb{E}(M_{\rm nsy})\Big\|_s\\ &\leq\underbrace{\Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}x_{2i}x_{3i}\boldsymbol{x}_{1i}\circ\boldsymbol{x}_{2i}\circ\boldsymbol{x}_{3i}-\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3|\mathcal{G}\Big)\Big\|_s}_{M_1\text{:main term}}\\ &+\underbrace{\Big\|\mathbb{E}\Big(x_1x_2x_3\boldsymbol{x}_1\circ\boldsymbol{x}_2\circ\boldsymbol{x}_3|\mathcal{G}\Big)-\boldsymbol{e}_1\circ\boldsymbol{e}_1\circ\boldsymbol{e}_1\Big\|_s}_{M_2\text{:bias term}},\end{aligned}$$

we will prove that M_2 is negligible in terms of convergence rate of M_1.

Bounding M_1. For simplicity, we define $\boldsymbol{x}_1'=\boldsymbol{x}_1|\mathcal{G},\ \boldsymbol{x}_2'=\boldsymbol{x}_2|\mathcal{G},\ \boldsymbol{x}_3'=\boldsymbol{x}_3|\mathcal{G}$, and $\{\boldsymbol{x}_{1i}',\boldsymbol{x}_{2i}',\boldsymbol{x}_{3i}'\}_{i=1}^n$ are n independent samples of $\{\boldsymbol{x}_1',\boldsymbol{x}_2',\boldsymbol{x}_3'\}$. According to the law of total probability, we have

$$\begin{aligned}&\mathbb{P}\Big(M_1\geq t\Big)\leq\mathbb{P}\Big(\mathcal{G}_n^c\Big)\\ &+\mathbb{P}\Big(\underbrace{\Big\|\frac{1}{n}\sum_{i=1}^n x_{1i}'\boldsymbol{x}_{1i}'\circ x_{3i}'\boldsymbol{x}_{2i}'\circ x_{i1}'\boldsymbol{x}_{3i}'-\mathbb{E}\Big(x_1'\boldsymbol{x}_1'\circ x_2'\boldsymbol{x}_2'\circ x_3'\boldsymbol{x}_3'\Big)\Big\|_s}_{M_{11}}\geq t\Big).\end{aligned}$$

According to Lemma 20, the entry of $x'_{1i}\boldsymbol{x}'_{1i}, x'_{2i}\boldsymbol{x}'_{2i}, x'_{3i}\boldsymbol{x}'_{3i}$ are sub-Gaussian random variable with ψ_2-norm M^2. Applying Lemma 18, 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

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

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

$$\mathbb{P}\Big(M_1 \geq \frac{64C_1}{C_2^3}\delta_{n,s}(\log n)^3\Big) \leq \frac{1}{p} + \frac{3e}{n^3}. \tag{S.10}$$

Bounding M_2. There exists $\varrho \in \mathbb{S}^{p-1}$ such that

$$M_2 = \Big|\mathbb{E}\Big(x_1x_2x_3(\boldsymbol{x}_1^\top\varrho)(\boldsymbol{x}_2^\top\varrho)(\boldsymbol{x}_3^\top\varrho)\Big|\mathcal{G}\Big) - (\boldsymbol{e}_1^\top\varrho)^3\Big|.$$

Since x_{1j} is independent of x_{1k} for any $j \neq k$, $\mathbb{E}(x_1(\boldsymbol{x}_1^\top\varrho)|\mathcal{G}) = \mathbb{E}(x_1^2\varrho_1|\mathcal{G})$. Then

$$\begin{aligned} M_2 &= \Big|\mathbb{E}\Big(x_1^2x_2^2x_3^2\varrho_1^3\Big|\mathcal{G}\Big) - \varrho_1^3\Big| \\ &= \Big|\varrho_1^3\mathbb{E}\Big(x_1^2\Big||x_1| \leq M\Big)\mathbb{E}\Big(x_2^2\Big||x_2| \leq M\Big)\mathbb{E}\Big(x_3^2\Big||x_3| \leq M\Big) - \varrho_1^3\Big|, \end{aligned}$$

where the second equation comes from the independence among each coordinate of $\{\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i}\}$.

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

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

Plugging them into M_2, we have

$$\begin{aligned} M_2 &\leq |\varrho_1^3|\Big|\Big(1 - 2Me^{-M^2/2}\Big)^3 - 1\Big| \\ &\leq \Big|12M^2e^{-M^2} - 6Me^{-M^2/2} - 8M^3e^{-3M^2/2}\Big| \\ &\leq \Big|26M^3e^{-M^2/2}\Big|, \end{aligned}$$

where the second inequality is due to $\|\varrho\|_2^2 = 1$ and the last inequality holds for a large $M > 0$. By the choice of $M = 2\sqrt{\log n/C_2}$, we have $M_2 \leq 208/C_2^{3/2}(\log n)^{\frac{3}{2}}/n^2$ for some constant C_2. When n is large, this rate is negligible comparing with (S.10)

Bounding M: We put the upper bounds of M_1 and M_2 together. After some adjustments for absolute constant, it suffices to obtain

$$\mathbb{P}\Big(M_1 + M_2 \leq C(\log n)^3\Big(\sqrt{\frac{s^3\log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big)\Big) \geq 1 - \frac{10}{n^3}.$$

This concludes the proof of non-symmetric part. The proof of symmetric part remains similar and thus is omitted here. ■

C Additional Proofs for Main Results

S.I 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 3.1, the original problem can be reformatted to a version of tensor denoising:

$$\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}, \quad \text{where} \quad \mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s). \tag{S.1}$$

The key difference between our model (S.1) and recent work is that $\mathcal{E}$ arises from empirical moment approximation, rather than the random observation noise considered in Anandkumar et al. (2014) and Sun et al. (2017). Next lemma gives an upper bound for the approximation error.

Lemma 9 (Approximation error of $\mathcal{T}_s$). Recall that $\mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s)$, where $\mathcal{T}_s$ is defined in (3.1). Suppose Condition 4 is satisfied and $s \le d \le Cs$. Then

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

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

Next we denote the following quantity for simplicity,

$$\gamma = C_2 \min \Big\{ \frac{R^{-1}}{6} - \frac{\sqrt{K}}{s}, \frac{R^{-1}}{4\sqrt{5}} - \frac{2}{\sqrt{s}} \Big(1 + \sqrt{\frac{K}{s}}\Big)^2 \Big\}, \tag{S.3}$$

where R is the singular value ratio, K is the CP-rank, s is the sparsity parameter, Γ is the incoherence parameter and C_2 is uniform constant.

Next lemma provides theoretical guarantees for sparse tensor decomposition method.

Lemma 10 . Suppose that the symmetric tensor denoising model (S.1) satisfies Conditions 1, 2 and 3 (i.e., the identifiability, parameter space and incoherence). Assume the number of initializations $L \ge K^{C_3 \gamma^{-4}}$ and the number of iterations $N \ge C_4 \log \Big(\gamma / \Big(\frac{1}{\eta^*_{\min}} \|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2\Big)\Big)$ for constants C_3, C_4, the truncation parameter $s \le d \le Cs$. Then the sparse-tensor-decomposition-based initialization satisfies

$$\max \Big\{ \|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*| \Big\} \le \frac{C_4}{\eta^*_{\min}} \|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2, \tag{S.4}$$

for any $k \in [K]$.

The proof of Lemma 10 essentially follows Theorem 3.9 in Sun et al. (2017), we thus omit the detailed proof here. The upper bound in (S.4) contains two terms: $\frac{C_4}{\eta^*_{\min}} \|\mathcal{E}\|_{s+d}$ and $\sqrt{K}\Gamma^2$, which are due to the empirical moment approximation and the incoherence among different $\boldsymbol{\beta}_k$, respectively.

Remark 4 . The guarantee of K-mean initialization scheme is hidden in Lemma 10 that provides a generic error bound for the sparse-tensor-decomposition-based initialization. Initialized by sparse SVD (Algorithm 3), we can prove that the K-means clustering outputs K cluster centers that are sufficiently close to the true components of the tensor.

Although the sparse tensor decomposition is not optimal in statistical rate, it does offer a reasonable initial estimation provided enough samples. Equipped with (S.2) and Condition 2, the right side of (S.4) reduces to

$$\begin{aligned} & \frac{C_4}{\eta^*_{\min}} \|\mathcal{E}\|_{s+d} + \sqrt{K}\Gamma^2 \\ \le \quad & 2C_1 C_4 K R \Big(\sqrt{\frac{s^3 \log^3(p/s)}{n^2}} + \sqrt{\frac{s \log(p/s)}{n}} \Big) (\log n)^4 + \sqrt{K}\Gamma^2, \end{aligned}$$

with probability at least $1-5/n$. Denote $C_0 = 4 \cdot 2160 \cdot C_1 C_4$. Using Conditions 3 and 5, we reach the conclusion that

$$\max\left\{\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*|\right\} \le K^{-1}R^{-2}/2160,$$

with probability at least $1-5/n$. ■

S.II 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 S.IV in the supplementary materials. For notation simplicity, we drop the superscript of $\eta_k^{(0)}$ in the following proof.

Lemma 11 . Let $t \ge 0$ be an integer. Suppose Conditions 1-5 hold and $\{\boldsymbol{\beta}_k^{(t)}, \eta_k\}$ satisfies the following upper bound

$$\sum_{k=1}^K \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \le 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2, \ \max_{k\in[K]} \left|\eta_k - \eta_k^*\right| \le \varepsilon_0, \tag{S.5}$$

with probability at least $1 - \mathcal{O}(K/n)$, where $\varepsilon_0 = K^{-1}R^{-\frac{4}{3}}/2160$. As long as the step size μ satisfies

$$0 < \mu \le \mu_0 = \frac{32R^{-20/3}}{3K[220+270K]^2}, \tag{S.6}$$

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

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

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

In order to apply Lemma 11, we prove that the required condition (S.5) holds at every iteration step t by induction. When $t = 0$, by (4.2) and Condition 2,

$$\left\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\right\|_2 \le \varepsilon_0, \ \left|\eta_k - \eta_k^*\right| \le \varepsilon_0, \ \text{for } k\in[K],$$

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

$$\begin{aligned}\left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2 &\le \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^{(0)} + \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2 \\ &\le |\sqrt[3]{\eta_k} - \sqrt[3]{\eta_k^*}| + \sqrt[3]{\eta_k^*}\left\|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\right\|_2.\end{aligned}$$

Note that

$$\left|\sqrt[3]{\eta_k} - \sqrt[3]{\eta_k^*}\right| \le \frac{\varepsilon_0}{(\sqrt[3]{\eta_k})^2 + \sqrt[3]{\eta_k\eta_k^*} + (\sqrt[3]{\eta_k^*})^2} \le \varepsilon_0\sqrt[3]{\eta_k^*}.$$

This implies

$$\left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(0)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2 \le 2\sqrt[3]{\eta_k^*}\varepsilon_0,$$

with probability at least $1-\mathcal{O}(1/n)$. Taking the summation over $k \in [K]$, we have

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

with probability at least $1-\mathcal{O}(K/n)$, which means (S.5) holds for $t=0$.

Suppose (S.5) holds at the iteration step $t-1$, which implies

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

Since Condition 5 automatically implies

$$\frac{n}{s\log p} \geq \frac{C_0\sigma^2R^{-\frac{2}{3}}\eta_{\min}^{*\frac{2}{3}}K}{64\varepsilon_0^2},$$

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, (S.5) holds at each iteration step.

Now we are able to use Lemma 11 recursively to complete the proof. Repeatedly using Lemma 11, we have for $t = 1, 2, \ldots,$

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

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

$$\sum_{k=1}^{K} \left\| \sqrt[3]{\eta_k}\boldsymbol{\beta}_k^{(T^*)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \right\|_2^2 \leq \frac{C_3\eta_{\min}^{*-\frac{4}{3}}}{8}\frac{\sigma^2 s\log p}{n}, \tag{S.7}$$

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

The next lemma shows that the Frobenius norm distance between two tensors can be bounded by the distances between each factors in their CP decomposition. The proof of this lemma is provided in Section S.V.

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 $|\eta_k - \eta_k^*| \le c$, then

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

Denote $\widehat{\mathscr{T}} = \sum_{k=1}^K \eta_k \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)}$. Combing (S.7) and Lemma 12, we have

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

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

S.III Proofs of Theorems 4: Minimax Lower Bounds

We first consider the proof of lower bound on a more general version of non-symmetric tensor estimation. Consider the class of incoherent sparse and low-rank tensors $\mathcal{F} = \{\mathscr{T} : \mathscr{T} = \sum_{k=1}^K \boldsymbol{\beta}_{1k} \circ \boldsymbol{\beta}_{2k} \circ \boldsymbol{\beta}_{3k}, \|\boldsymbol{\beta}_{i,k}\|_0 \le s \text{ for } i = 1,2,3, k = 1,\dots,K\}$ and the measurement tensor can be written as $\mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i$. Without loss of generality we assume $p = \max\{p_1, p_2, p_3\}$. We uniformly randomly generate $\{\Omega^{(k,m)}\}_{\substack{m=1,\dots,M \\ k=1,\dots,K}}$ as MK subsets of $\{1,\dots,p\}$ with cardinality of s. Here $M > 0$ is a large integer to be specified later. Then we construct $\{\boldsymbol{\beta}^{(k,m)}\}_{\substack{m=1,\dots,M \\ k=1,\dots,K}} \subseteq \mathbb{R}^p$ as

$$\boldsymbol{\beta}_j^{(k,m)} = \begin{cases} \sqrt{\lambda}, & \text{if } j \in \Omega^{(k,m)}; \\ 0, & \text{if } j \notin \Omega^{(k,m)}. \end{cases}$$

$\lambda > 0$ will also be specified a little while later. Clearly, $\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\|_2^2 \le 2s\lambda$ for any $1 \le k \le K$, $1 \le m_1, m_2 \le M$. Additionally, $|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}|$ satisfies the hyper-geometric distribution: $\mathbb{P}\left(\left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right| = t\right) = \frac{\binom{s}{t}\binom{p-s}{s-t}}{\binom{p}{s}}$.

Let $w^{(k,m_1,m_2)} = \left|\Omega^{(k,m_1)} \cap \Omega^{(k,m_2)}\right|$, then for any $s/2 \le t \le s$,

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

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

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

By setting $\eta = \log((p-s+1)/(8s))$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} \geq \frac{3sK}{4}\right) = \mathbb{P}\left(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2} \geq \frac{sK}{4}\right) \\
\leq & \frac{\mathbb{E}\exp\left(\eta(\sum_{k=1}^{K} w^{(k,m_1,m_2)} - \frac{sK}{2})\right)}{\exp\left(\eta \cdot \frac{sK}{4}\right)} = \frac{\prod_{k=1}^{K} \mathbb{E}\exp\left(\eta(w^{(k,m_1,m_2)} - \frac{s}{2})\right)}{\exp(\eta \cdot \frac{sK}{4})} \\
\leq & \left(1 + (4s/(p-s+1))^{s/2} \cdot 2\right)^K \exp\left(-\frac{sK}{4}\log\left(\frac{p-s+1}{8s}\right)\right) \\
\leq & \exp\left(-c_0 sK \log(p/s)\right)
\end{aligned}$$

for some small uniform constant $c_0 > 0$.

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

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

then we further have

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

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

$$\begin{aligned}
\frac{sK\lambda}{2} &\leq \min_{1\leq m_1<m_2\leq M} \sum_{k=1}^{K} \left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \\
&\leq \max_{1\leq m_1<m_2\leq M} \sum_{k=1}^{K} \left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2 \leq 2sK\lambda.
\end{aligned} \tag{S.8}$$

For the rest of the proof, we fix $\left\{\boldsymbol{\beta}^{(k,m)}\right\}_{\substack{k=1,\ldots,K \\ m=1,\ldots,M}}$ to be the set of vectors satisfying (S.8).

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

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

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 $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2), \quad i=1,\dots,n$. Clearly, $\left(\boldsymbol{y}^{(m)}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}\right)$ follows a joint distribution, which may vary based on different values of m.

In this step, we analyze the Kullback-Leibler divergence between different distribution pairs:

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

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,

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

Therefore, for any $m_1 \neq m_2$,

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

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

$$\begin{aligned}
\|\mathscr{T}^{(m_1)} - \mathscr{T}^{(m_2)}\|_F &= \left\|\sum_{k=1}^{K}(\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)})\circ\boldsymbol{e}^{(k)}\circ\boldsymbol{e}^{(k)}\right\|_F\\
&=\sqrt{\sum_{k=1}^{K}\left\|\boldsymbol{\beta}^{(k,m_1)} - \boldsymbol{\beta}^{(k,m_2)}\right\|_2^2} \overset{\text{(S.8)}}{\ge} \sqrt{\frac{sK\lambda}{2}}.
\end{aligned}$$

By generalized Fano's Lemma (see, e.g., Yu (1997)),

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

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

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

which has finished the proof of non-symmetric tensor estimation model.

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,

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

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

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

The rest of the proof essentially follows from the proof of non-asymmetric tensor estimation model. ■

S.IV Proof of Theorem 5: High-order Stein's Lemma

The proof of this theorem follows from the one of Theorem 6 in Janzamin et al. (2014). For the sake of completeness, we restate the detail here. Applying the recursion representation of score function (S.4), 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 24) 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. ■

D Proofs of Several Lemmas

S.I Proofs of Lemmas 3, and 4: Moment Calculation

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

S.I.1 Proof of Lemma 2

By the definition of $\{y_i\}$ in (S.1) & (S.2), we have

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

First, we observe $\mathbb{E}(\epsilon_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i) = 0$ due to the independence between ϵ_i and $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}$. Then, we consider a single component from a single observation

$$M = \mathbb{E}((\boldsymbol{\beta}_{1k}^{*\top}\boldsymbol{u}_i)(\boldsymbol{\beta}_{2k}^{*\top}\boldsymbol{v}_i)(\boldsymbol{\beta}_{3k}^{*\top}\boldsymbol{w}_i)\boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i),\ i \in [n], k \in [K].$$

For notation simplicity, we drop the subscript i for i-th observation and k for k-th component such that

$$M = \mathbb{E}\Big((\boldsymbol{\beta}_1^{*\top}\boldsymbol{u})(\boldsymbol{\beta}_2^{*\top}\boldsymbol{v})(\boldsymbol{\beta}_3^{*\top}\boldsymbol{w})\boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\Big) \in \mathbb{R}^{p_1 \times p_2 \times p_3}. \tag{S.2}$$

Each entry of M can be calculated as follows

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

which implies $M = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \boldsymbol{\beta}_3$. Combining with n observations and K components, we can obtain

$$\mathbb{E}(\mathcal{T}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\eta_k^* \boldsymbol{\beta}_{1k} \circ \boldsymbol{\beta}_{2k} \circ \boldsymbol{\beta}_{3k}.$$

This finished our proof. ■

S.I.2 Proof of Lemma 3

In this subsection, we provide an alternative and more direct proof for Lemma 3. We consider a similar single component with a symmetric structure, namely, $M_s = \mathbb{E}\Big((\boldsymbol{\beta}^{*\top}\boldsymbol{x})^3 \boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x}\Big)$. Based on the symmetry of both underlying tensor and sketchings, we will verify the following three cases:

- When $i = j = k$, then

$$\begin{aligned}
M_{s_{iii}} &= \mathbb{E}\Big(\beta_i^* x_i + \sum_{m \neq i} \beta_m^* x_m\Big)^3 x_i^3 \\
&= \mathbb{E}\Big(\beta_i^{*3} x_i^3 + 3\beta_i^{*2} x_i^2 \big(\sum_{m \neq i} \beta_m^* x_m\big) \\
&\quad + 3\beta_i^* x_i \big(\sum_{m \neq i} \beta_m^* x_m\big)^2 + \big(\sum_{m \neq i} \beta_m^* x_m\big)^3\Big) x_i^3 \\
&= 15\beta_i^{*3} + 9\beta_i^* \sum_{m \neq i} \beta_m^{*2} = 9\beta_i^* + 6\beta_i^{*3}.
\end{aligned}$$

  The last equation is due to $\|\boldsymbol{\beta}^*\|_2 = 1$.

- When $i \neq j \neq k$, then

$$\begin{aligned}
M_{s_{ijk}} &= \mathbb{E}\Big(\beta_i^* x_i + \beta_j^* x_j + \beta_k^* x_k + \sum_{m \neq i,j,k} \beta_m^* x_m\Big)^3 x_i x_j x_k \\
&= \mathbb{E}\big(\beta_i^* x_i + \beta_j^* x_j + \beta_k^* x_k\big)^3 x_i x_j x_k \\
&= 6\beta_i^* \beta_j^* \beta_k^*.
\end{aligned}$$

- When $i = j \neq k$, then

$$\begin{aligned}
M_{s_{iik}} &= \mathbb{E}\Big(\beta_i^* x_i + \beta_k^* x_k + \sum_{m \neq i,k} \beta_m^* x_m\Big)^3 x_i^2 x_k \\
&= 9\beta_i^{*2}\beta_k^* + 3\beta_k^{*3} + 3\beta_k^* \big(\sum_{m \neq i,k} \beta_m^{*2}\big) \\
&= 9\beta_i^{*2}\beta_k^* + 3\beta_k^* \big(\sum_{m \neq i} \beta_m^{*2}\big) \\
&= 3\beta_k^* + 6\beta_i^{*2}\beta_k^*.
\end{aligned}$$

Therefore, it is sufficient to calculate M_s by

$$\begin{aligned}
M_s =& 3\sum_{k=1}^K \eta_k^* \Big(\sum_{m=1}^p \boldsymbol{\beta}_k^* \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ \boldsymbol{\beta}_k^*\Big) \\
&+ 6\sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*.
\end{aligned}$$

The first term is the bias term due to correlations among symmetric sketchings. Denote $M_1 = \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i$ and note that $\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i\Big) = 3\sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_k^*$. Therefore, the empirical first-order moment M_1 could be used to remove the bias term as follows

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

This finishes our proof. ■

S.I.3 Proof of Lemma 4

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 3. For the subscript involving 0, the following two calculations work for removing the bias,

$$\begin{aligned}
\mathbb{E}\Big(\frac{1}{3}\mathcal{T}_s - \frac{1}{6}(\sum_{k=1}^{p}\mathcal{T}_{s,[k,k,i]} - (p+1)\boldsymbol{a}_i)\Big) &= \beta_0^2\beta_i. \\
\mathbb{E}\Big(\frac{1}{2p-2}(\sum_{k=1}^{p}\mathcal{T}_{s[0,k,k]} - (p+2)\mathcal{T}_{s[0,0,0]})\Big) &= \beta_0^3.
\end{aligned}$$

This ends the proof. ■

S.II Proof of Lemma 5

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 Ledoux and Talagrand (2013)), we have

$$\Big\|\sum_{i=1}^{n} a_i X_i\Big\|_p \leq 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i X_i\Big\|_p = 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i |X_i|\Big\|_p, \tag{S.3}$$

where $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables and we notice that $\varepsilon_i X_i$ and $\varepsilon_i|X_i|$ are identically distributed. Moreover, if $|X_i| \geq \beta$, the definition of Z_i implies that $|X_i| = Z_i + \beta$. And if $|X_i| < \beta$, we have $Z_i = 0$. Thus, we have $|X_i| \leq Z_i + \beta$ at any time and it leads to

$$2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i |X_i|\Big\|_p \quad \leq \quad 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i(\beta + Z_i)\Big\|_p. \tag{S.4}$$

By triangle inequality,

$$2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i(\beta+Z_i)\Big\|_p \le 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i Z_i\Big\|_p + 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i\beta\Big\|_p. \tag{S.5}$$

Next, we will bound the second term of the RHS of (S.5). In particular, we will utilize Khinchin-Kahane inequality, whose formal statement is included in Lemma 25 for the sake of completeness. From Lemma 25 we have

$$\begin{aligned}\Big\|\sum_{i=1}^{n} a_i\varepsilon_i\beta\Big\|_p &\le \Big(\frac{p-1}{2-1}\Big)^{1/2}\Big\|\sum_{i=1}^{n} a_i\varepsilon_i\beta\Big\|_2\\ &\le \beta\sqrt{p}\Big\|\sum_{i=1}^{n} a_i\varepsilon_i\Big\|_2. \end{aligned}\tag{S.6}$$

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

$$\begin{aligned}\Big(\mathbb{E}\Big(\sum_{i=1}^{n}\varepsilon_i a_i\Big)^2\Big)^{1/2} &= \Big(\mathbb{E}\Big(\sum_{i=1}^{n}\varepsilon_i^2 a_i^2 + 2\sum_{1\le i<j\le n}\varepsilon_i\varepsilon_j a_i a_j\Big)\Big)^{1/2}\\ &= \Big(\sum_{i=1}^{n} a_i^2\mathbb{E}\varepsilon_i^2 + 2\sum_{1\le i<j\le n} a_i a_j\mathbb{E}\varepsilon_i\mathbb{E}\varepsilon_j\Big)^{1/2}\\ &= \Big(\sum_{i=1}^{n} a_i^2\Big)^{1/2} = \|\boldsymbol{a}\|_2. \end{aligned}\tag{S.7}$$

Combining inequalities (S.4)-(S.7),

$$2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i|X_i|\Big\|_p \le 2\Big\|\sum_{i=1}^{n} a_i\varepsilon_i Z_i\Big\|_p + 2\beta\sqrt{p}\|\boldsymbol{a}\|_2. \tag{S.8}$$

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

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

which implies

$$\Big\|\sum_{i=1}^{n} a_i\varepsilon_i Z_i\Big\|_p \le \Big\|\sum_{i=1}^{n} a_i\varepsilon_i Y_i\Big\|_p = \Big\|\sum_{i=1}^{n} a_i Y_i\Big\|_p, \tag{S.9}$$

since $\varepsilon_i Y_i$ and Y_i have the same distribution due to symmetry. Combining (S.8) and (S.9) together, we reach

$$\Big\|\sum_{i=1}^{n} a_i X_i\Big\|_p \le 2\beta\sqrt{p}\|\boldsymbol{a}\|_2 + 2\Big\|\sum_{i=1}^{n} a_i Y_i\Big\|_p. \tag{S.10}$$

For $0<\alpha<1$, it follows Lemma 23 that

$$\Big\|\sum_{i=1}^{n} a_i Y_i\Big\|_p \le C_1(\alpha)(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_\infty), \tag{S.11}$$

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

For $\alpha \geq 1$, we will combine Lemma 22 and the method of the integration by parts to pass from tail bound result to moment bound result. Recall that for every non-negative random variable X, integration by parts yields the identity

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)dt.$$

Applying this to $X = |\sum_{i=1}^n a_i Y_i|^p$ and changing the variable $t = t^p$, then we have

$$\begin{aligned}
\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p &= \int_0^\infty \mathbb{P}\Big(|\sum_{i=1}^n a_i Y_i| \geq t\Big) p t^{p-1} dt \\
&\leq \int_0^\infty 2\exp\Big(-c\min\Big(\frac{t^2}{\|\boldsymbol{a}\|_2^2}, \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big) p t^{p-1} dt, \qquad \text{(S.12)}
\end{aligned}$$

where the inequality is from Lemma 22 for all $p \geq 2$ and $1/\alpha + 1/\alpha^* = 1$. In this following, we bound the integral in three steps:

1. If $\frac{t^2}{\|\boldsymbol{a}\|_2^2} \leq \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (S.12) 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,

$$\Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \frac{p^{1/p}}{(2c)^{1/2}}\sqrt{p}\|\boldsymbol{a}\|_2 \leq \frac{\sqrt{2}}{\sqrt{c}}\sqrt{p}\|\boldsymbol{a}\|_2. \qquad \text{(S.13)}$$

2. If $\frac{t^2}{\|\boldsymbol{a}\|_2^2} > \frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}$, (S.12) reduces to

$$\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p \leq 2p\int_0^\infty \exp\Big(-c\frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big) t^{p-1} dt.$$

   Letting $t' = ct^\alpha/\|\boldsymbol{a}\|_{\alpha^*}^\alpha$, we have

$$\begin{aligned}
2p\int_0^\infty \exp\Big(-c\frac{t^\alpha}{\|\boldsymbol{a}\|_{\alpha^*}^\alpha}\Big)\Big) t^{p-1} dt &= \frac{2p\|\boldsymbol{a}\|_{\alpha^*}^p}{\alpha c^{p/\alpha}}\int_0^\infty e^{-t'} t'^{p/\alpha-1} dt' \\
&= \frac{2}{\alpha}\frac{p\|\boldsymbol{a}\|_{\alpha^*}^p}{c^{p/\alpha}}\Gamma(\frac{p}{\alpha}) \leq \frac{2}{\alpha}\frac{p\|\boldsymbol{a}\|_{\alpha^*}^p}{c^{p/\alpha}}(\frac{p}{\alpha})^{p/\alpha}.
\end{aligned}$$

   Thus,

$$\Big(\mathbb{E}|\sum_{i=1}^n a_i Y_i|^p\Big)^{\frac{1}{p}} \leq \frac{2p^{1/p}}{(c\alpha)^{1/\alpha}} p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*} \leq \frac{4}{(c\alpha)^{1/\alpha}} p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}. \qquad \text{(S.14)}$$

3. Overall, we have the following by combining (S.13) and (S.14),

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

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

$$\Big\| \sum_{i=1}^{n} a_i Y_i \Big\|_p \le C_2(\alpha)\Big(\sqrt{p}\|\boldsymbol{a}\|_2 + p^{1/\alpha}\|\boldsymbol{a}\|_{\alpha^*}\Big). \tag{S.15}$$

Since $0 < \beta < 1$, the conclusion can be reached by combining (S.10),(S.11) and (S.15). ■

S.III Proof of Lemma 9

Firstly, let us consider the non-symmetric perturbation error analysis using model (S.1). According to Lemma 2, the exact form of $\mathcal{E} = \mathcal{T} - \mathbb{E}(\mathcal{T})$ is given by

$$\mathcal{E} = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i - \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*.$$

We decompose it by a concentration term ($\mathcal{E}_1$) and a noise term ($\mathcal{E}_2$) as follows,

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

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

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

By using Lemma 1 and $s \le d \le Cs$, it suffices to have for some absolute constant C_{11},

$$\|\mathcal{E}_{1k}\|_{s+d} \le C_{11}\delta_{n,p,s}, \text{ where } \delta_{n,p,s} = (\log n)^3\Big(\sqrt{\frac{s^3 \log^3(p/s)}{n^2}} + \sqrt{\frac{s\log(p/s)}{n}}\Big),$$

with probability at least $1 - 10/n^3$, where $\|\cdot\|_{s+d}$ is the sparse tensor spectral norm defined in (2.3). Equipped with the triangle inequality, the sparse tensor spectral norm for $\mathcal{E}_1$ can be bounded by

$$\|\mathcal{E}_1\|_{s+d} \le C_{11}\delta_{n,p,s}\sum_{k=1}^{K} \eta_k^*, \tag{S.16}$$

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 18, we have for some absolute constant C_{12}

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

with probability at least $1 - 1/p$. According to Lemma 21, we have

$$\mathbb{P}\Big(\|\mathcal{E}_2\|_{s+d} \ge C_{12}\sigma \log n \delta_{n,p,s}\Big) \le \frac{1}{p} + \frac{3}{n} \le \frac{4}{n}. \tag{S.17}$$

Bounding $\mathcal{E}$: Putting (S.16) and (S.17) together, we obtain

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

with probability at least $1-5/n$. Under Condition 4, we have

$$\|\mathcal{E}\|_{s+d} \le 2C_1\sum_{k=1}^{K}\eta_k^*\delta_{n,p,s}\log n,$$

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. ■

S.IV Proof of Lemma 11

Lemma 11 quantifies one step update for thresholded gradient update. The proof consists of two parts. First, we evaluate an oracle estimator $\{\widetilde{\boldsymbol{\beta}}_k^{(t+1)}\}_{k=1}^K$ with known support information, which is defined as

$$\widetilde{\boldsymbol{\beta}}_k^{(t+1)} = \varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k^{(t)})}\Big(\boldsymbol{\beta}_k^{(t)} - \frac{\mu}{\phi}\nabla_k\mathcal{L}(\boldsymbol{\beta}_k^{(t)})_{F^{(t)}}\Big). \tag{S.18}$$

Here,

- $h(\boldsymbol{\beta}_k^{(t)})$ is the k-th component of $h(\boldsymbol{B}^{(t)})$ defined in (3.2).
- $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}) = (\nabla_1\mathcal{L}(\boldsymbol{\beta}_1),\cdots,\nabla_K\mathcal{L}(\boldsymbol{\beta}_K))$.
- $F^{(t)} = \cup_{k=1}^K F_k^{(t)}$, where $F_k^{(t)} = \mathrm{supp}(\boldsymbol{\beta}_k^*)\cup\mathrm{supp}(\boldsymbol{\beta}_k^{(t)})$.
- For a vector $\boldsymbol{x}\in\mathbb{R}^p$ and a subset $A\subset\{1,\dots,p\}$, 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 (S.5) is satisfied and $|F|\lesssim Ks$. As long as the step size $\mu\le 32R^{-20/3}/(3K[220+270K]^2)$, we obtain the upper bound for $\{\widetilde{\boldsymbol{\beta}}_k^+\}$,

$$\begin{aligned}\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \le &\Big(1-32\mu\frac{R^{-\frac{8}{3}}}{K^2}\Big)\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2\\ &+ 2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n},\end{aligned} \tag{S.19}$$

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

The proof of Lemma 13 is postponed to the Section S.VI. Next lemma guarantees that with high probability, $\{\boldsymbol{\beta}_k^+\}_{k=1}^K$ is equivalent to the oracle update $\{\widetilde{\boldsymbol{\beta}}_k^+\}_{k=1}^K$ with high probability.

Lemma 14 . Recall that the truncation level $h(\boldsymbol{\beta}_k)$ is defined as

$$h(\boldsymbol{\beta}_k) = \frac{\sqrt{4\log np}}{n}\sqrt{\sum_{i=1}^{n}\Big(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3 - y_i\Big)^2\Big(\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2\Big)^2}. \tag{S.20}$$

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 S.VI. 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 $1-(21K^2+11K+4Ks)/n$,

$$\begin{aligned}\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \le & \Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^{K}\Big\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\Big\|_2^2 \\ & + 2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n},\end{aligned} \tag{S.21}$$

This ends the proof. ■

S.V Proof of Lemma 12

We consider a more general setting that the tensor is not necessary to be symmetric such that

$$\mathscr{T} = \sum_{k=1}^{K}\eta_k\boldsymbol{\beta}_k\circ\boldsymbol{\beta}_k\circ\boldsymbol{\beta}_k, \mathscr{T}^* = \sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*\circ\boldsymbol{\beta}_k^*.$$

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

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

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

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

Since $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we have

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

Equipped with Cauchy-Schwarz inequality, RHS can be further bounded by

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

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

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

For the non-symmetric tensor estimation model, we have

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

Following the same strategy above, we obtain

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

This ends the proof. ■

S.VI Proof of Lemma 13

First of all, let us state a lemma to illustrate the effect of weight ϕ.

Lemma 15 . Consider $\{y_i\}_{i=1}^n$ come from either non-symmetric tensor estimation model (S.1) or symmetric tensor estimation model (3.1). Suppose Conditions 3-5 hold. Then $\phi=\frac{1}{n}\sum_{i=1}^n y_i^2$ is upper and lower bounded by

$$(16-6\Gamma^3-9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2 \leq \frac{1}{n}\sum_{i=1}^{n}y_i^2 \leq (16+6\Gamma^3+9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2,$$

with probability at least $1-(K^2+K+3)/n$, where Γ is the incoherence parameter.

According to Lemma 15, $\frac{1}{n}\sum_{i=1}^n y_i^2$ approximates $(\sum_{k=1}^K \eta_k^*)^2$ up to some constants with high probability. Moreover, we know that from (S.5), $\max_k |\eta_k - \eta_k^*| \le \varepsilon_0$ for some small ε_0. Based on those two facts described above, we replace η_k by η_k^* and ϕ by $(\sum_{k=1}^K \eta_k^*)^2$ for the sake of completeness. Note that this change could only result in some constant scale changes for final results. Similar simplification was used in matrix recovery scenario (Tu et al., 2015). Therefore, we define the weighted estimator and weighted true parameter as $\bar{\boldsymbol{\beta}}_k = \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k$, $\bar{\boldsymbol{\beta}}_k^* = \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*$. Correspondingly, define the gradient function $\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)$ on F as

$$\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)_F = \frac{6\sqrt[3]{\eta_k^*}}{n}\sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'})^3 - y_i\Big)(\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2 \boldsymbol{x}_{i_F},$$

and its noiseless version as

$$\nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)_F = \frac{6\sqrt[3]{\eta_k^*}}{n}\sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'})^3 - \sum_{k'=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3\Big)(\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2 \boldsymbol{x}_{i_F}. \tag{S.22}$$

According to the definition of thresholding function in Section 3.2, $\widetilde{\boldsymbol{\beta}}_k^+$ 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 $\text{supp}(\boldsymbol{\gamma}_k) \subset F$, $\|\boldsymbol{\gamma}_k\|_\infty \le 1$ and $h(\bar{\boldsymbol{\beta}}_k)$ is defined as

$$h(\bar{\boldsymbol{\beta}}_k) = \frac{\sqrt{4\log(np)}}{n}\sqrt{\sum_{i=1}^n \Big(\sum_{k=1}^K (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^3 - y_i\Big)^2 {\eta_k^*}^{\frac{2}{3}} (\boldsymbol{x}_{i_F}^\top \bar{\boldsymbol{\beta}}_k)^2}. \tag{S.23}$$

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:

$$\begin{aligned}
&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k^*}\widetilde{\boldsymbol{\beta}}_k^+ - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\
&= \sum_{k=1}^K \Big\| \boldsymbol{z}_k - \frac{\mu\sqrt[3]{\eta_k^*}}{\phi}\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k) + \frac{\mu\sqrt[3]{\eta_k^*}}{\phi} h(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\gamma}_k \Big\|_2^2 \\
&= \underbrace{\sum_{k=1}^K \Big\| \boldsymbol{z}_k - \mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k) \Big\|_2^2}_{\text{A: gradient update effect}} + \underbrace{\sum_{k=1}^K \Big\| \frac{\mu\sqrt[3]{\eta_k^*}}{\phi} h(\bar{\boldsymbol{\beta}}_k)\gamma_k \Big\|_2^2}_{\text{B: threshoding effect}} \\
&+ \underbrace{\sum_{k=1}^K \Big\langle \boldsymbol{z}_k - \mu\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k), \frac{\mu\sqrt[3]{\eta_k}}{\phi} h(\bar{\boldsymbol{\beta}}_k)\gamma_k \Big\rangle}_{\text{C: cross term}}.
\end{aligned} \tag{S.24}$$

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

S.VI.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

$$\begin{aligned}
A &= \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2 - 2\mu \sum_{k=1}^{K} \Big\langle \frac{\sqrt[3]{\eta_k^*}}{\phi} \nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{z}_k \Big\rangle + \mu^2 \sum_{k=1}^{K} \Big\| \frac{\sqrt[3]{\eta_k^*}}{\phi} \nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k) \Big\|_2^2 \\
&\leq \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_2^2 - 2\mu \underbrace{\sum_{k=1}^{K} \Big\langle \frac{\sqrt[3]{\eta_k^*}}{\phi} \nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k), \boldsymbol{z}_k \Big\rangle}_{A_1} + 2\mu^2 \underbrace{\sum_{k=1}^{K} \Big\| \frac{\sqrt[3]{\eta_k^*}}{\phi} \nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k) \Big\|_2^2}_{A_2} \\
&\quad + 2\mu^2 \underbrace{\sum_{k=1}^{K} \Big\| \frac{\sqrt[3]{\eta_k^*}}{\phi} \Big(\nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k) - \nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big) \Big\|_2^2}_{A_3} \\
&\quad + 2\mu \underbrace{\sum_{k=1}^{K} \Big\langle \boldsymbol{z}_k, \frac{\sqrt[3]{\eta_k^*}}{\phi} \Big(\nabla_k \widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k) - \nabla_k \mathcal{L}(\bar{\boldsymbol{\beta}}_k)\Big) \Big\rangle}_{A_4},
\end{aligned} \tag{S.25}$$

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 (Candès et al., 2015).

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

$$K^{-2} R^{-\frac{2}{3}} \eta_{\max}^{*-\frac{4}{3}} \leq \frac{(\sqrt[3]{\eta_k^*})^2}{\phi} = \frac{(\sqrt[3]{\eta_k^*})^2}{(\sum_{k=1}^{K} \eta_k^*)^2} \leq K^{-2} R^{\frac{2}{3}} \eta_{\min}^{*-\frac{4}{3}}. \tag{S.26}$$

According to the definition of noiseless gradient $\nabla_k \widetilde{\mathcal{L}}(\boldsymbol{\beta}_k)$ and $\boldsymbol{z}_k$, A_1 can be expanded and decomposed sequentially by nine terms,

$$
\begin{aligned}
A_1 \geq & K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}\Big[\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{11} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{12} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^3\Big) \Leftarrow A_{13} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{14} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{15} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{16} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{17} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{18} \\
& + \frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^3\Big)\Big] \Leftarrow A_{19},
\end{aligned}
\tag{S.27}
$$

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 S.1: high-order Gaussian moment;

2. Argue that the empirical version is concentrated around their expectation with high probability through Lemma 8: 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 S.1, the expectation of A_{11} can be calculated explicitly as

$$
\begin{aligned}
\mathbb{E}(A_{11}) = & \, 36\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)^2(\boldsymbol{z}_{k'}^\top\boldsymbol{z}_k) \Leftarrow I_1 \\
& + 72\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_{k'}^*) \Leftarrow I_2 \\
& + 108\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_k^\top\bar{\boldsymbol{\beta}}_k^*) \Leftarrow I_3 \\
& + 54\sum_{k=1}^{K}\sum_{k'=1}^{K}(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k'}^*)(\bar{\boldsymbol{\beta}}_k^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top\boldsymbol{z}_k) \Leftarrow I_4.
\end{aligned}
\tag{S.28}
$$

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

$$I_1 \geq 36\eta_{\min}^{*4/3}\Big[\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2 - \Gamma^2\Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2\Big]$$

$$I_2 \geq 72\eta_{\min}^{*4/3}\Big[\sum_{k=1}^K (\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_k^*)^2 - \Gamma\Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2\Big]$$

$$I_3 \geq 108\eta_{\min}^{*4/3}\Big[\sum_{k=1}^K (\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_k^*)^2 - \Gamma\Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2\Big]$$

$$I_4 \geq 54\eta_{\min}^{*4/3}\Big\|\sum_{k=1}^K \boldsymbol{z}_k\Big\|_2^2 \geq 0,$$

where Γ is the incoherence parameter. Putting the above four bounds together, they jointly provide

$$\mathbb{E}(A_{11}) \geq 36\eta_{\min}^{*4/3}\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2 - \Big(36\eta_{\min}^{*4/3}\Gamma^2 + 180\eta_{\min}^{*4/3}\Gamma\Big)\Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2. \tag{S.29}$$

On the other hand, repeatedly using Lemma 8, 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,

$$\Big|A_{11} - \mathbb{E}(A_{11})\Big| \leq 18C\frac{(\log n)^3}{\sqrt{n}}(\sqrt[3]{\eta_{\max}^*})^4\Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2, \tag{S.30}$$

with probability at least $1-K^2/n$. Combining (S.29) and (S.30), we obtain with probability at least $1-K^2/n$,

$$\begin{aligned}&K^{-2}R^{-\frac{2}{3}}\eta_{\max}^{*-\frac{4}{3}}A_{11} \\ \geq&\Big[36K^{-2}R^{-\frac{8}{3}} - K^{-\frac{3}{2}}\Big(216R^{-\frac{8}{3}}\Gamma + 18C\frac{(\log n)^3}{\sqrt{n}}\Big)\Big]\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2,\end{aligned} \tag{S.31}$$

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 S.1, the expectation of A_{12} can be calculated as

$$\begin{aligned}\mathbb{E}(A_{12}) \quad = \quad & 36\sum_{k=1}^K\sum_{k'=1}^K (\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_{k'}^{*\top})^2(\boldsymbol{z}_{k'}^\top \bar{\boldsymbol{\beta}}_k^*) \Leftarrow \; I_1 \\ & +72\sum_{k=1}^K\sum_{k'=1}^K (\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_{k'}^*)(\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_{k'}^\top \boldsymbol{z}_k) \Leftarrow \; I_2 \\ & +108\sum_{k=1}^K\sum_{k'=1}^K (\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_{k'})(\boldsymbol{z}_{k'}^\top \bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_k^\top \bar{\boldsymbol{\beta}}_k^*) \Leftarrow \; I_3 \\ & +54\sum_{k=1}^K\sum_{k'=1}^K (\bar{\boldsymbol{\beta}}_{k'}^{*\top}\bar{\boldsymbol{\beta}}_{k'}^*)(\boldsymbol{z}_{k'}^\top \bar{\boldsymbol{\beta}}_k^*)(\boldsymbol{z}_k^\top \boldsymbol{z}_k) \Leftarrow \; I_4.\end{aligned}$$

Let us analyze I_1 first. Under (S.5), $\|\boldsymbol{z}_k\|_2 \le \varepsilon_0 \sqrt[3]{\eta_k^*}$, it suffices to show that

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

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

$$\mathbb{E}(A_{12}) \ge -270 \eta_{\max}^{*\frac{4}{3}} \varepsilon_0 \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2. \tag{S.32}$$

By Lemma 8, we obtain for some absolute constant C,

$$\begin{aligned}&K^{-2} R^{-\frac{2}{3}} \eta_{\max}^{*-\frac{4}{3}} A_{12} \\ \ge & K^{-2} R^{-\frac{2}{3}} \eta_{\max}^{*-\frac{4}{3}} \Big[\mathbb{E}(A_{12}) - 18C \eta_{\max}^{*\frac{4}{3}} \varepsilon_0 \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2\Big)^2 \frac{(\log n)^3}{\sqrt{n}}\Big] \\ \ge & -K^{-1} R^{-\frac{2}{3}} \varepsilon_0 \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big),\end{aligned} \tag{S.33}$$

with probability at least $1 - K^2/n$. The detail derivation is the same as in (S.31), so we omit here.

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

$$\begin{aligned}&K^{-\frac{1}{2}} \eta_{\max}^{*-\frac{4}{3}} A_{14} \ge -K^{\frac{1}{2}} \varepsilon_0 \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big) \\ &K^{-\frac{1}{2}} \eta_{\max}^{*-\frac{4}{3}} A_{13}, A_{15}, A_{17} \ge -K^{\frac{1}{2}} \varepsilon_0^2 \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big) \\ &K^{-\frac{1}{2}} \eta_{\max}^{*-\frac{4}{3}} A_{16}, A_{18} \ge -K^{\frac{1}{2}} \varepsilon_0^3 \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big) \\ &K^{-\frac{1}{2}} \eta_{\max}^{*-\frac{4}{3}} A_{19} \ge -K^{\frac{1}{2}} \varepsilon_0^4 \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big).\end{aligned} \tag{S.34}$$

Putting (S.31), (S.33) and (S.34) together, we have with probability at least $1 - 9K^2/n$,

$$\begin{aligned}A_1 \quad \ge \quad & \Big[36 K^{-2} R^{-\frac{8}{3}} - K^{-\frac{3}{2}} \Big(2160 R^{-\frac{3}{3}} \Gamma + 18C \frac{(\log n)^3}{\sqrt{n}}\Big) \\ & -8\varepsilon_0 K^{-1} R^{-\frac{2}{3}} \Big(270 + 18C \frac{(\log n)^3}{\sqrt{n}}\Big)\Big] \Big(\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big).\end{aligned}$$

For the above bound,

- When the sample size satisfies $n \ge (18 C K^{1/2} R^{8/3} (\log n)^3)^2$, we have

$$\max\Big\{18 K^{-\frac{3}{2}} C \frac{(\log n)^3}{\sqrt{n}}, 8\varepsilon_0 K^{-1} R^{-\frac{2}{3}} 18C \frac{(\log n)^3}{\sqrt{n}}\Big\} \le K^{-2} R^{-\frac{8}{3}}.$$

- When $\varepsilon_0 \le K^{-1} R^{-2}/2160$, we have

$$8\varepsilon_0 K^{-1} R^{-\frac{2}{3}} 270 \le K^{-2} R^{-\frac{8}{3}}.$$

- 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 (S.5). Thus, we are able to simplify A_1 by

$$A_1 \geq 32K^{-2}R^{-\frac{8}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big), \tag{S.35}$$

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

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

$$\begin{aligned}A_2 &= \sum_{k=1}^{K}\Big\|\frac{1}{\phi}\sqrt[3]{\eta_k^*}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k)\Big\|_2^2\\ &= \sup_{\boldsymbol{w}\in\mathbb{S}^{Ks-1}}\Big|\Big\langle\sum_{k=1}^{K}\frac{\sqrt[3]{\eta_k^*}}{\phi}\nabla_k\widetilde{\mathcal{L}}(\bar{\boldsymbol{\beta}}_k),\boldsymbol{w}\Big\rangle\Big|^2,\end{aligned} \tag{S.36}$$

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 (S.22), A_2' is explicitly written as

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

Following by (S.26) and (S.27), similar decomposition can be made for A_2' as follows, where the only difference is that we replace one $\boldsymbol{x}_i^\top\boldsymbol{z}_k$ by $\boldsymbol{x}_i^\top\boldsymbol{w}$.

$$\begin{aligned}A_2' \leq\ & K^{-2}R^{\frac{2}{3}}\eta_{\min}^{*-\frac{4}{3}}\Big[\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{21}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{22}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})^2\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\boldsymbol{w})\Big) \Leftarrow A_{23}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{24}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{25}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_{k'})\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{26}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)^2\Big) \Leftarrow A_{27}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}2(\boldsymbol{x}_i^\top\boldsymbol{z}_k)(\boldsymbol{x}_i^\top\boldsymbol{w})(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k^*)\Big) \Leftarrow A_{28}'\\ &+\frac{6}{n}\sum_{i=1}^{n}\Big(\sum_{k'=1}^{K}3(\boldsymbol{x}_i^\top\boldsymbol{z}_{k'})^3\sum_{k=1}^{K}(\boldsymbol{x}_i^\top\boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top\boldsymbol{w})\Big)\Big]. \Leftarrow A_{29}'\end{aligned}$$

Let's bound A'_{21} first. By using the same technique when calculating $\mathbb{E}(A_{11})$ in (S.28), we derive an upper bound for $\mathbb{E}(A'_{21})$,

$$
\begin{aligned}
\mathbb{E}(A'_{21}) \quad &\leq \quad 36\eta_{\max}^{*\frac{4}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2 + (K-1)\sum_{k=1}^{K}\Gamma\|\boldsymbol{z}_k\|_2\Big) \\
&+ \quad 180\eta_{\max}^{*\frac{4}{3}}\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2 + (K-1)\sum_{k=1}^{K}\Gamma\|\boldsymbol{z}_k\|_2\Big) + 54\eta_{\max}^{*\frac{4}{3}}\Big(K\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big).
\end{aligned}
$$

Equipped with Lemma 1 and the definition of tensor spectral norm (2.3), it suffices to bound A'_{21} by

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

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

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

$$
\begin{aligned}
A'_2 \quad &\leq \quad K^{-2}R^2\Big[216 + 270K + 18CK\delta_{n,p,s}\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big) \\
&\leq \quad K^{-2}R^2\Big[220 + 270K\Big]\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2\Big),
\end{aligned}
$$

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

$$
A_2 \leq K^{-1}R^4[220 + 270K]^2\Big(\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2\Big), \tag{S.37}
$$

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

$$
\begin{aligned}
A_3 \quad &= \quad \sum_{k=1}^{K}\Big\|\frac{(\sqrt[3]{\eta_k^*})^2}{\phi}\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2\boldsymbol{x}_i\Big\|_2^2 \\
&\leq \quad K^{-4}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\sum_{k=1}^{K}\Big(\sqrt{Ks}\max_j\frac{6}{n}\sum_{i=1}^{n}\epsilon_i(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^2x_{ij}\Big)^2,
\end{aligned}
$$

where the second inequality comes from (S.26). For fixed $\{\epsilon_i\}_{i=1}^n$, applying Lemma 8, we have

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

with probability at least $1-1/n$. Together with Lemma 21, 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 σ is the noise level. According to (S.5),

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

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 $j\in[Ks]$,

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

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

$$A_3 \leq C\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}\sigma^2K^{-2}\frac{s(\log n)^3}{n}, \tag{S.38}$$

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 $1-4K/n$,

$$A_4 \leq 2C\sigma\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}K^{-1}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{2}{3}}. \tag{S.39}$$

This term is negligible in terms of the order when comparing with (S.38).

Summary. Putting the bounds (S.35), (S.37), (S.38) and (S.39) together, we achieve an upper bound for gradient update effect as follows,

$$\begin{aligned} A \leq &\Big(1-64\mu K^{-2}R^{-\frac{8}{3}}+2\mu^2K^{-1}R^4[220+270K]^2\Big)\sum_{k=1}^{K}\|\boldsymbol{z}_k\|_2^2 \\ &+4\mu CK^{-2}\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}\frac{\sigma^2s(\log n)^3}{n}, \end{aligned} \tag{S.40}$$

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

S.VI.2 Bounding thresholding effect

The thresholding effect term in (S.24) can also be decomposed into optimization error and statistical error. Recall that B can be explicitly written as

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

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

$$
\begin{aligned}
B \le& \mu^2 \frac{64Ks\log p}{n}\Big[\underbrace{\frac{1}{n}\sum_{i=1}^n \Big(\sum_{k'=1}^K (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^3 - \sum_{k'=1}^K (\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^3\Big)\Big(\sum_{k=1}^K \frac{\eta_k^{*\frac{4}{3}}}{\phi^2}(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4\Big)}_{B_1\text{:optimization error}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \sum_{k=1}^K \frac{\eta_k^{*\frac{4}{3}}}{\phi^2}(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k)^4}_{B_2\text{:statistical error}}\Big].
\end{aligned}
$$

Bounding B_1. This optimization error term shares similar structure with (S.36) but with higher order. Therefore, we follow the same idea as we did in bounding (S.36). Following by (S.26) and some basic expansions and inequalities,

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

The main term is $(\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_k^*)^4$ according to the order of $\bar{\boldsymbol{\beta}}_k^*$. We bound the main term first. Note that there exists some positive large constant C such that

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

Together with Lemma 8 and (S.5), we have

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{k'=1}^K \Big(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{z}_k)^2(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'}^*)^4(\boldsymbol{x}_i^\top \bar{\boldsymbol{\beta}}_{k'})^4\Big) \\
\le& C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)K^2\eta_{\max}^{*\frac{8}{3}}(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2.
\end{aligned}
$$

with probability at least $1-3K^2/n$. Overall, the upper bound of B_1 takes the form

$$
\begin{aligned}
B_1 \le& K^{-2}R^{\frac{4}{3}}\eta_{\min}^{*-\frac{8}{3}}\Big[18C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)K^2\eta_{\max}^{*\frac{8}{3}}(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2\Big] \\
\le& R^4 18C\Big(1+\frac{(\log n)^5}{\sqrt{n}}\Big)(1+K^{\frac{1}{2}}\varepsilon_0)^4\sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2,
\end{aligned}
\tag{S.41}
$$

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 8, we have

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

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

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

From Lemma 21, with probability at least $1-3/n$,

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

Combining the above two inequalities, we obtain

$$\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\boldsymbol{x}_i^\top\bar{\boldsymbol{\beta}}_k)^4\Big| \le 6C_0\sigma^2\|\bar{\boldsymbol{\beta}}_k\|_2^4, \tag{S.42}$$

with probability at least $1-7/n$. Plugging in the definition of ϕ and (S.5), B_2 is upper bounded by

$$B_2 \le 6C_0\sigma^2(1+K^{\frac{1}{2}}\varepsilon_0)^4\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}K^{-3}, \tag{S.43}$$

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

Summary. Putting the bounds (S.41) and (S.43) together, we have similar upper bound for thresholded effect,

$$B \le C_2\mu^2R^4\sum_{k=1}^K\|\boldsymbol{z}_k\|_2^2 + C_3\mu^2\eta_{\min}^{*-\frac{4}{3}}R^{\frac{8}{3}}K^{-2}\frac{\sigma^2 s\log p}{n}, \tag{S.44}$$

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

S.VI.3 Ensemble

From the definition of γ_k, it's not hard to see actually the cross term C is equal to zero. Combining the upper bound of gradient update effect (S.40) and thresholding effect (S.44) 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 μ satisfies

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

we reach the conclusion

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\widetilde{\boldsymbol{\beta}}_k^{+}-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
\le&\Big(1-32\mu K^{-2}R^{-\frac{8}{3}}\Big)\sum_{k=1}^{K}\left\|\sqrt[3]{\eta_k}\boldsymbol{\beta}_k-\sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^*\right\|_2^2\\
&+2C_3\mu^2R^{-\frac{8}{3}}\eta_{\min}^{*-\frac{4}{3}}\frac{\sigma^2K^{-2}s\log p}{n},
\end{aligned}\tag{S.45}$$

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

S.VII Proof of Lemma 14

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

$$\frac{\partial}{\partial\beta_{kj}}\mathcal{L}(\boldsymbol{\beta}_k)=\frac{2}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{K}\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^3-y_i\Big)\eta_k(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)^2x_{ij},\tag{S.46}$$

and it's not hard to see the independence between $\{\boldsymbol{x}_i^\top\boldsymbol{\beta}_k,y_i\}$ and x_{ij}. Applying standard Hoeffding's inequality, we have with probability at least $1-\frac{1}{n^2p^2}$,

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

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,

$$\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)\Big)=\varphi_{\frac{\mu}{\phi}h(\boldsymbol{\beta}_k)}\Big(\boldsymbol{\beta}_k-\frac{\mu}{\phi}\nabla_{\boldsymbol{\beta}_k}\mathcal{L}(\boldsymbol{\beta}_k)_F\Big),\tag{S.47}$$

holds for $k\in[K]$, with probability at least $1-\frac{1}{n^2p}$. (S.47) 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. ■

S.VIII Proof of Lemma 15

First, we consider symmetric case. According to the definition of $\{y_i\}_{i=1}^n$ from symmetric tensor estimation model (3.1), we separate the random noise ϵ_i by the following expansion,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}y_i^2&=\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3+\epsilon_i\Big]^2\\
&=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2}_{I_1}+\underbrace{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3}_{I_2}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}_{I_3}.
\end{aligned}\tag{S.48}$$

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

$$\begin{aligned}&(\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2\\&=\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6+2\sum_{k_i<k_j}\eta_{k_i}^*\eta_{k_j}^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3.\end{aligned}\tag{S.49}$$

As shown in Corollary S.1, the expectations of above two parts takes forms of

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

Recall that $\|\boldsymbol{\beta}_k^*\|_2=1$ for any $k\in[K]$ and Condition 3 implies for any $k_i\neq k_j$, $|\boldsymbol{\beta}_{k_i}^{*\top}\boldsymbol{\beta}_{k_j}^*|\le\Gamma$, where Γ is the incoherence parameter. Thus, $\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3$ is upper bounded by

$$\Big|\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3\Big|\le 6\Gamma^3+9\Gamma,\ \text{for any } k_i\neq k_j.\tag{S.50}$$

By using the concentration result in Lemma 8, we have with probability at least $1-1/n$

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6-\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^6)\Big|\le C_1\frac{(\log n)^3}{\sqrt{n}},\\&\Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3-\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_i}^*)^3(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{k_j}^*)^3)\Big|\le C_1\frac{(\log n)^3}{\sqrt{n}}.\end{aligned}\tag{S.51}$$

Putting (S.49),(S.50) and (S.51) together, this essentially provides an upper bound for I_1, namely

$$\frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3)^2\le\Big(15+6\Gamma^3+9\Gamma+2C_1\frac{(\log n)^3}{\sqrt{n}}\Big)(\sum_{k=1}^{K}\eta_k^*)^2,\tag{S.52}$$

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 8, we have

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

This further implies that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\eta_k^*(\boldsymbol{x}_i^\top\boldsymbol{\beta}_k^*)^3\epsilon_i\le(\sum_{k=1}^{K}\eta_k^*)C_2\frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}\sigma,\tag{S.53}$$

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

Bounding I_3. As shown in Lemma 21, the random noise ϵ_i with sub-exponential tail satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \le C_3\sigma^2. \tag{S.54}$$

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

Overall, putting (S.52), (S.53) and (S.54) together, we have with probability at least $1-(K^2+4K+3)/n$,

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

Under Conditions 4 & 5, the above bound reduces to

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 \le (16+6\Gamma^3+9\Gamma)(\sum_{k=1}^{K}\eta_k^*)^2,$$

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

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

■

E Matrix Form Gradient and Stochastic Gradient descent

S.I Matrix Formulation of Gradient

In this section, we provide detail derivations for (3.5).

Lemma S.1. Let $\boldsymbol{\eta}=(\eta_1,\dots,\eta_K)\in\mathbb{R}^{K\times 1}$, $\boldsymbol{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)\in\mathbb{R}^{p\times n}$ and $\boldsymbol{B}=(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K)\in\mathbb{R}^{p\times K}$. The gradient of symmetric tensor estimation empirical risk function (3.3) can be written in a matrix form as follows

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

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

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

Correspondingly, each part can be written as a matrix form,

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

This implies that $[((\boldsymbol{B}^\top\boldsymbol{X})^\top)^3\eta-\boldsymbol{y}]^\top[(((\boldsymbol{B}^\top\boldsymbol{X})^\top)^2\odot\eta^\top)^\top\odot\boldsymbol{X}]^\top\in\mathbb{R}^{1\times pK}$. Note that $\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B},\boldsymbol{\eta})=(\nabla\mathcal{L}_1(\boldsymbol{\beta}_1)^\top,\dots,\nabla\mathcal{L}_K(\boldsymbol{\beta}_K)^\top)\in\mathbb{R}^{1\times pK}$. The conclusion can be easily derived. ■

S.II Stochastic Gradient descent

Stochastic thresholded gradient descent is a stochastic approximation of the gradient descent optimization method. Note that the empirical risk function (3.3) that can be written as a sum of differentiable functions. Followed by (3.5), the gradient of (3.3) evaluated at i-th sketching $\{y_i, \boldsymbol{x}_i\}$ can be written as

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

Thus, the overall gradient $\nabla_{\boldsymbol{B}}\mathcal{L}_i(\boldsymbol{B},\boldsymbol{\eta})$ defined in (3.5) 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

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

F Technical Lemmas

Lemma 16 . Suppose $\boldsymbol{x} \in \mathbb{R}^p$ is a standard Gaussian random vector. For any non-random vector $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^p$, we have the following tensor expectation calculation,

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

where $\boldsymbol{e}_m$ is a canonical vector in $\mathbb{R}^p$.

Proof. Recall that for a standard Gaussian random variable x, its odd moments are zero and even moments are $\mathbb{E}(x^6) = 15, \mathbb{E}(x^4) = 4$. Expanding the LHS of (S.1) and comparing LHS and RHS, we will reach the conclusion. Details are omitted here. ∎

Lemma 17 . Suppose $\boldsymbol{u} \in \mathbb{R}^{p_1}, \boldsymbol{v} \in \mathbb{R}^{p_2}, \boldsymbol{w} \in \mathbb{R}^{p_3}$ are independent standard Gaussian random vectors. For any non-random vector $\boldsymbol{a} \in \mathbb{R}^{p_1}, \boldsymbol{b} \in \mathbb{R}^{p_2}, \boldsymbol{c} \in \mathbb{R}^{p_3}$, we have the following tensor expectation calculation

$$\mathbb{E}\Big((\boldsymbol{a}^\top \boldsymbol{u})(\boldsymbol{b}^\top \boldsymbol{v})(\boldsymbol{c}^\top \boldsymbol{w})\boldsymbol{u}\circ\boldsymbol{v}\circ\boldsymbol{w}\Big) = \boldsymbol{a}\circ\boldsymbol{b}\circ\boldsymbol{c}. \tag{S.2}$$

Proof. Due to the independence among $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$, the conclusion is easy to obtain by using the moment of standard Gaussian random variable. ∎

Note that in the left side of (S.1), it involves an expectation of rank-one tensor. When multiplying any non-random rank-one tensor with same dimensionality, i.e. $\boldsymbol{a}_1\circ\boldsymbol{b}_1\circ\boldsymbol{c}_1$, on both sides, it will facilitate us to calculate the expectation of product of Gaussian vectors, see next Lemma for details.

Lemma S.1. Suppose $\boldsymbol{x} \in \mathbb{R}^p$ is a standard Gaussian random vector. For any non-random vector $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d} \in$

$\mathbb{R}^p$, we have the following expectation calculation

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^6 = 15\|\boldsymbol{a}\|_2^6,\\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^5(\boldsymbol{x}^\top \boldsymbol{b}) = 15\|\boldsymbol{a}\|_2^4(\boldsymbol{a}^\top \boldsymbol{b}),\\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^4(\boldsymbol{x}^\top \boldsymbol{b})^2 = 12\|\boldsymbol{a}\|_2^2(\boldsymbol{a}^\top \boldsymbol{b})^2 + 3\|\boldsymbol{a}\|_2^4\|\boldsymbol{b}\|_2^2,\\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^3 = 6(\boldsymbol{a}^\top \boldsymbol{b})^3 + 9(\boldsymbol{a}^\top \boldsymbol{b})\|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2,\\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^2(\boldsymbol{x}^\top \boldsymbol{c}) = 6(\boldsymbol{a}^\top \boldsymbol{b})^2(\boldsymbol{a}^\top \boldsymbol{c}) + 6(\boldsymbol{a}^\top \boldsymbol{b})(\boldsymbol{b}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{a})\\
&+3(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{b}^\top \boldsymbol{b})(\boldsymbol{a}^\top \boldsymbol{a}),\\
&\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{a})^2(\boldsymbol{x}^\top \boldsymbol{b})(\boldsymbol{x}^\top \boldsymbol{c})^2(\boldsymbol{x}^\top \boldsymbol{d}) = 2(\boldsymbol{a}^\top \boldsymbol{c})^2(\boldsymbol{b}^\top \boldsymbol{d}) + 4(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{b}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{d})\\
&+6(\boldsymbol{a}^\top \boldsymbol{c})(\boldsymbol{a}^\top \boldsymbol{b})(\boldsymbol{c}^\top \boldsymbol{d}) + 3(\boldsymbol{c}^\top \boldsymbol{x})(\boldsymbol{b}^\top \boldsymbol{d})(\boldsymbol{a}^\top \boldsymbol{a}).
\end{aligned}$$

Proof. Note that $\mathbb{E}((\boldsymbol{x}^\top \boldsymbol{a})^3(\boldsymbol{x}^\top \boldsymbol{b})^3) = \mathbb{E}((\boldsymbol{x}^\top \boldsymbol{a})^3\langle \boldsymbol{x}\circ\boldsymbol{x}\circ\boldsymbol{x}, \boldsymbol{b}\circ\boldsymbol{b}\circ\boldsymbol{b}\rangle)$. Then we can apply the general result in Lemma 16. 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 18 . Suppose $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_n^\top)^\top, \boldsymbol{Y} = (\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_n^\top)^\top, \boldsymbol{Z} = (\boldsymbol{z}_1^\top, \cdots, \boldsymbol{z}_n^\top)^\top$ are three $n\times p$ random matrices. The ψ_2-norm of each entry is bounded, s.t. $\|X_{ij}\|_{\psi_2} = K_x, \|Y_{ij}\|_{\psi_2} = K_y, \|Z_{ij}\|_{\psi_2} = K_z$. We assume the row of $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ are independent. There exists an absolute constant C such that,

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

Here, $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (2.3) and $\delta_{n,p,s} = \sqrt{s\log(ep/s)/n} + \sqrt{s^3\log(ep/s)^3/n^2}$.

Proof. Bounding spectral norm always relies on the construction of the ϵ-net. Since we will bound a sparse tensor spectral norm, our strategy is to discrete the sparse set and construct the ϵ-net on each one. Let us define a sparse set $\mathcal{B}_0 = \{\boldsymbol{x}\in\mathbb{R}^p, \|\boldsymbol{x}\|_2 = 1, \|\boldsymbol{x}\|_0 \leq s\}$. And let $\mathcal{B}_{0,s}$ be the s-dimensional set defined by $\mathcal{B}_{0,s} = \{\boldsymbol{x}\in\mathbb{R}^s, \|\boldsymbol{x}\|_2 = 1\}$. Note that $\mathcal{B}_0$ is corresponding to s-sparse unit vector set which can be expressed as a union of subsets of dimension s by expanding some zeros, namely $\mathcal{B}_0 = \cup\,\mathcal{B}_{0,s}$. There should be at most $\binom{p}{s} \leq (\frac{ep}{s})^s$ such set $\mathcal{B}_{0,s}$.

Recalling the definition of sparse tensor spectral norm in (2.3), we have

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

Instead of constructing the ϵ-net on $\mathcal{B}_0$, we will construct an ϵ-net for each of subsets $\mathcal{B}_{0,s}$. Define $\mathcal{N}_{\mathcal{B}_{0,s}}$ as the 1/2-set of $\mathcal{B}_{0,s}$. From Lemma 3.18 in Ledoux (2005), the cardinality of $\mathcal{N}_{0,s}$ is bounded by 5^s. By Lemma 19, we obtain

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

By rotation invariance of sub-Gaussian random variable, $\langle \boldsymbol{x}_i, \boldsymbol{\chi}_1\rangle$, $\langle \boldsymbol{y}_i, \boldsymbol{\chi}_2\rangle$, $\langle \boldsymbol{z}_i, \boldsymbol{\chi}_3\rangle$ are still sub-Gaussian random variables with ψ_2-norm bounded by K_x, K_y, K_z, respectively. Applying Lemma 8 and union bound over $\mathcal{N}_{\mathcal{B}_{0,s}}$, the right hand side of (S.3) can be bounded by

$$\mathbb{P}\Big(\text{RHS} \geq 8K_xK_yK_zC\Big(\sqrt{\frac{\log \delta^{-1}}{n}} + \sqrt{\frac{(\log \delta^{-1})^3}{n^2}}\Big)\Big) \leq (5^s)^3\delta,$$

for any $0 < \delta < 1$.

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

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

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 19 (Tensor Covering Number(Lemma 4 in Nguyen et al. (2015))). Let $\mathbb{N}$ be an ϵ-net for a set $\boldsymbol{B}$ associated with a norm $\|\cdot\|$. Then, the spectral norm of a d-mode tensor $\mathcal{A}$ is bounded by

$$\begin{aligned} & \sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{d-1}\in\boldsymbol{B}} \|\mathcal{A}\times_1 \boldsymbol{x}_1 \ldots \times_{d-1} \boldsymbol{x}_{d-1}\|_2 \\ \leq & \Big(\frac{1}{1-\varepsilon}\Big)^{d-1} \sup_{\boldsymbol{x}_1\cdots\boldsymbol{x}_{d-1}\in\mathbb{N}} \|\mathcal{A}\times_1 \boldsymbol{x}_1 \cdots \times_{d-1} \boldsymbol{x}_{d-1}\|_2. \end{aligned}$$

This immediately implies that the spectral norm of a d-mode tensor $\mathcal{A}$ is bounded by

$$\|\mathcal{A}\|_2 \leq (\frac{1}{1-\epsilon})^{d-1} \sup_{\boldsymbol{x}_1\ldots\boldsymbol{x}_{d-1}\in\mathcal{N}} \|\mathcal{A}\times_1 \boldsymbol{x}_1 \cdots \times_{d-1} \boldsymbol{x}_{d-1}\|_2,$$

where $\mathbb{N}$ is the ϵ-net for the unit sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$.

Lemma 20 (Sub-Gaussianess of the Product of Random Variables). Suppose X_1 is a bounded random variable with $|X_1| \leq K_1$ almost surely for some K_1 and X_2 is a sub-Gaussian random variable with Orlicz norm $\|X_2\|_{\psi_2}K_2$. Then X_1X_2 is still a sub-Gaussian random variable with Orlicz norm $\|X_1X_2\|_{\psi_2} = K_1K_2$.

Proof: Following the definition of sub-Gaussian random variable, we have

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

holds for all $t \geq 0$. This ends the proof. ■

Lemma 21 (Tail Probability for the Sum of Sub-exponential Random Variables (Lemma A.7 in Cai et al. (2016))). 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

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\Big| \le C_0\sigma\sqrt{\frac{\log n}{n}},\ \|\boldsymbol{\epsilon}\|_\infty \le C_0\sigma\log n,$$
$$\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\Big| \le C_0\sigma^2, \Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^4\Big| \le C_0\sigma^4,$$

for some constant C_0.

Lemma 22 (Tail Probability for the Sum of Weibull Distributions (Lemma 3.6 in Adamczak et al. (2011))). Let $\alpha \in [1, 2]$ and $Y_1, \dots, Y_n$ be independent symmetric random variables satisfying $\mathbb{P}(|Y_i| \ge t) = \exp(-t^\alpha)$. Then for every vector $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$ and every $t \ge 0$,

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

Proof. It is a combination of Corollaries 2.9 and 2.10 in Talagrand (1994).

Lemma 23 (Moments for the Sum of Weibull Distributions (Corollary 1.2 in Bogucki (2015))). Let $X_1, X_2, \dots, X_n$ be a sequence of independent symmetric random variables satisfying $\mathbb{P}(|Y_i| \ge t) = \exp(-t^\alpha)$, where $0 < \alpha < 1$. Then, for $p \ge 2$ and some constant $C(\alpha)$ which depends only on α,

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

Lemma 24 (Stein's Lemma (Stein et al., 2004)). 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 25 (Khinchin-Kahane Inequality (Theorem 1.3.1 in De la Pena and Giné (2012))). Let $\{a_i\}_{i=1}^n$ a finite non-random sequence, $\{\varepsilon_i\}_{i=1}^n$ be a sequence of independent Rademacher variables and $1 < p < q < \infty$. Then

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

Lemma 26 . Suppose each non-zero element of $\{\boldsymbol{x}_k\}_{k=1}^K$ is drawn from standard Gaussian distribution and $\|\boldsymbol{x}_k\|_0 \le s$ for $k \in [K]$. Then we have for any $0 < \delta \le 1$,

$$\mathbb{P}\Big(\max_{1\le k_1<k_2\le K} |\langle \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}\rangle| \le C\sqrt{s}\sqrt{\log K + \log 1/\delta}\Big) \ge 1 - \delta,$$

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 $x_{k_1i} \neq 0$ and $x_{k_2j} \neq 0$. From the definition of $\mathcal{S}_{k_1k_2}$, we know that $|\mathcal{S}_{k_1k_2}| \le s$ and $\boldsymbol{x}_{k_1}^\top \boldsymbol{x}_{k_2} = \sum_{j=1}^p x_{k_1j}x_{k_2j} = \sum_{j\in\mathcal{S}_{k_1k_2}} x_{k_1j}x_{k_2j}$. We apply standard Hoeffding's concentration inequality,

$$\mathbb{P}\Big(|\langle \boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}\rangle| \ge t\Big) = \mathbb{P}\Big(|\sum_{j\in\mathcal{S}_{k_1k_2}} x_{k_1j}x_{k_2j}| \ge t\Big) \le e\exp\Big(-\frac{ct^2}{s}\Big).$$

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

References

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88.

Anandkumar, A., Ge, R., and Janzamin, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bogucki, R. (2015). Suprema of canonical weibull processes. *Statistics & Probability Letters*, 107:253–263.

Cai, T. T., Li, X., and Ma, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251.

Cai, T. T. and Zhang, A. (2015). Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138.

Caiafa, C. F. and Cichocki, A. (2013). Multidimensional compressed sensing and their applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(6):355–380.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Chen, H., Raskutti, G., and Yuan, M. (2016). Non-convex projected gradient descent for generalized low-rank tensor regression. *arXiv preprint arXiv:1611.10349*.

Chen, Y., Chi, Y., and Goldsmith, A. J. (2015). Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059.

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*.

Friedland, S., Li, Q., and Schonfeld, D. (2014). Compressive sensing of sparse tensors. *IEEE Transactions on Image Processing*, 23(10):4438–4447.

Friedland, S. and Lim, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Håstad, J. (1990). Tensor rank is np-complete. *Journal of algorithms (Print)*, 11(4):644–654.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45.

Hitczenko, P., Montgomery-Smith, S., and Oleszkiewicz, K. (1997). Moment inequalities for sums of certain independent symmetric random variables. *Studia Math*, 123(1):15–42.

Hung, H., Lin, Y.-T., Chen, P., Wang, C.-C., Huang, S.-Y., and Tzeng, J.-Y. (2016). Detection of genegene interactions using multistage sparse and low-rank regression. *Biometrics*, 72(1):85–94.

Janzamin, M., Sedghi, H., and Anandkumar, A. (2014). Score function features for discriminative learning: matrix and tensor framework. *arXiv preprint arXiv:1412.2863*.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998.

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51:455–500.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. Wiley Series in Probability and Statistics.

Ledoux, M. (2005). *The concentration of measure phenomenon*. Number 89. American Mathematical Soc.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, pages 1–16.

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

Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545.

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

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Montanari, A. and Sun, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 73–81, Bejing, China. PMLR.

Nguyen, N. H., Drineas, P., and Tran, T. D. (2015). Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229.

Raskutti, G., Yuan, M., and Chen, H. (2018). Convex regularization for high-dimensional multi-response tensor regression. *The Annals of Statistics*, to appear.

Rauhut, H., Schneider, R., and Stojanac, Ž. (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262.

Richard, E. and Montanari, A. (2014). A statistical model for tensor pca. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2897–2905. Curran Associates, Inc.

Romera-Paredes, B., Aung, M. H., Bianchi-Berthouze, N., and Pontil, M. (2013). Multilinear multitask learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1444–III–1452. JMLR.org.

Sidiropoulos, N. D. and Kyrillidis, A. (2012). Multi-way compressed sensing for sparse low-rank tensors. *IEEE Signal Processing Letters*, 19(11):757–760.

Stein, C., Diaconis, P., Holmes, S., Reinert, G., et al. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pages 1–25. Institute of Mathematical Statistics.

Sun, W. W. and Li, L. (2017). Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944.

Sun, W. W., Lu, J., Liu, H., and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916.

Talagrand, M. (1994). The supremum of some canonical processes. *American Journal of Mathematics*, 116(2):283–325.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2015). Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*.

Vershynin, R. (2012). *Compressed sensing*, chapter Introduction to the non-asymptotic analysis of random matrices, pages 210–268. Cambridge Univ. Press.

Wang, Z., Liu, H., and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of statistics*, 42(6):2164.

Yu, B. (1997). Assouad, fano, and le cam. *Festschrift for Lucien Le Cam*, 423:435.

Yuan, M. and Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068.

Yuan, M. and Zhang, C.-H. (2017). Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766.

Zhang, A. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964.

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

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2016). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552.