Communication-Efficient Estimation of

Regularized Smoothed Support Tensor Machine

Zihao Song[1], Lei Wang[2], Riquan Zhang[3], Weihua Zhao[1]

1. Nantong University, 2. Nankai University

3.Shanghai University of International Business and Economics

Supplementary Material

S1 Background on t-product

Before presenting the t-product and related theories (Kilmer et al., 2008; Kilmer and Martin, 2011), we introduce some additional notations. For a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, denote the block circular matrix $\mathrm{bcirc}(\boldsymbol{\mathcal{A}})$ as

$$\overline{\boldsymbol{A}} = \mathrm{bcirc}(\boldsymbol{\mathcal{A}}) := \begin{bmatrix} \boldsymbol{\mathcal{A}}^{(1)} & \boldsymbol{\mathcal{A}}^{(I_3)} & \cdots & \boldsymbol{\mathcal{A}}^{(2)} \\ \boldsymbol{\mathcal{A}}^{(2)} & \boldsymbol{\mathcal{A}}^{(1)} & \cdots & \boldsymbol{\mathcal{A}}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\mathcal{A}}^{(I_3)} & \boldsymbol{\mathcal{A}}^{(I_3-1)} & \cdots & \boldsymbol{\mathcal{A}}^{(1)} \end{bmatrix} \in \mathbb{R}^{I_1 I_3 \times I_2 I_3}.$$

The block vector-unfolding operator $\mathrm{bvec}(\cdot)$ and its inverse operator $\mathrm{bvfold}(\cdot)$

are defined as, respectively,

$$\mathrm{bvec}(\boldsymbol{\mathcal{A}}) := [\boldsymbol{\mathcal{A}}^{(1)}; \cdots ; \boldsymbol{\mathcal{A}}^{(I_3)}], \ \ \mathrm{bvfold}(\mathrm{bvec}(\boldsymbol{\mathcal{A}})) := \boldsymbol{\mathcal{A}}.$$

Definition S1.1 *(t-product): The t-product between* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times r \times I_3}$ *and* $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{r \times I_2 \times I_3}$ *is defined as*

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} := \mathrm{bvfold}(\mathrm{bcirc}(\boldsymbol{\mathcal{A}}) \cdot \mathrm{bvec}(\boldsymbol{\mathcal{B}})) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$$

where $*$ *indicate the t-product.*

To obtain the faster computational efficiency, the discrete Fourier transformation (DFT), connecting t-product and standard matrix product, is introduced. Denote $\boldsymbol{F}_n = [\omega_1, \omega_2, \cdots, \omega_n]$ as the DFT matrix, where

$$\omega_i = [\vartheta^{0\times(i-1)}; \vartheta^{1\times(i-1)}; \cdots ; \vartheta^{(n-1)\times(i-1)}],$$

$\vartheta = e^{-\frac{2\pi\varrho}{n}}$ and $\varrho = \sqrt{-1}$. Thus, $\boldsymbol{F}_n$ is a unitary matrix, that is, $\boldsymbol{F}_n^{\mathsf{H}}\boldsymbol{F}_n = \boldsymbol{F}_n\boldsymbol{F}_n^{\mathsf{H}} = n\boldsymbol{I}_n$ and $\boldsymbol{F}_n^{-1} = \frac{1}{n}\boldsymbol{F}_n^{\mathsf{H}}$. Then, the block-circulant matrices can be

blockdiagonalized, i.e.,

$$(\boldsymbol{F}_{I_3} \otimes \boldsymbol{I}_{I_1})\overline{\boldsymbol{A}}(\boldsymbol{F}_{I_3}^{\mathsf{H}} \otimes \boldsymbol{I}_{I_2}) = \widetilde{\boldsymbol{A}} = \begin{bmatrix} \widetilde{\boldsymbol{\mathcal{A}}}^{(1)} & & & \\ & \widetilde{\boldsymbol{\mathcal{A}}}^{(2)} & & \\ & & \ddots & \\ & & & \widetilde{\boldsymbol{\mathcal{A}}}^{(I_3)} \end{bmatrix} \tag{S1.1}$$

where $\widetilde{\boldsymbol{\mathcal{A}}}$ is obtained by $\widetilde{\boldsymbol{\mathcal{A}}}(i,j,:) = \boldsymbol{F}_{I_3}\boldsymbol{\mathcal{A}}(i,j,:)$. By the unitarity and (S1.1),

$$\|\boldsymbol{\mathcal{A}}\|_F = \frac{1}{\sqrt{I_3}}\|\widetilde{\boldsymbol{A}}\|_F,\ \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \frac{1}{I_3}\langle \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{A}} \rangle,$$
$$\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} \Leftrightarrow \widetilde{\boldsymbol{\mathcal{C}}} = \widetilde{\boldsymbol{\mathcal{A}}} \cdot \widetilde{\boldsymbol{\mathcal{B}}}.$$

Next, we introduce some other definitions on tensor, generalized from the matrix case.

Definition S1.2 *(Conjugate transpose, Lu et al. (2018)): The conjugate transpose of $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ is represented as $\boldsymbol{\mathcal{A}}^{\mathsf{H}} \in \mathbb{C}^{I_2 \times I_1 \times I_3}$, achieved by transposing each frontal slice with conjugation and subsequently reversing the sequence of transposed frontal slices from 2 to I_3.*

Definition S1.3 *(Orthogonal tensor): A tensor $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is orthogonal if $\boldsymbol{\mathcal{Q}} * \boldsymbol{\mathcal{Q}}^{\mathsf{H}} = \boldsymbol{\mathcal{Q}}^{\mathsf{H}} * \boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{I}}$.*

Definition S1.4 *(F-diagonal tensor): A tensor is considered f-diagonal when*

each of its frontal slices represents a diagonal matrix.

Definition S1.5 *(Identity tensor): A tensor $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{I\times I\times I_3}$ is called as identity tensor if its first frontal slice is the identity matrix and others are all zero.*

Definition S1.6 *(t-SVD): For a given $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times I_3}$, it has the following factorization,*

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}},$$

where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{I_1\times I_1\times I_3}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{I_2\times I_2\times I_3}$ are orthogonal tensors, and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{I_1\times I_2\times I_3}$ is a f-diagonal tensor.

Note that the entries of the first frontal slice in $\boldsymbol{\mathcal{S}}$ possess the decreasing property, i.e.,

$$\boldsymbol{\mathcal{S}}(1,1,1) \geq \boldsymbol{\mathcal{S}}(2,2,1) \geq \cdots \geq \boldsymbol{\mathcal{S}}(I_1 \wedge I_2, I_1 \wedge I_2, 1). \tag{S1.2}$$

The property (S1.2) holds because the inverse DFT gives,

$$\boldsymbol{\mathcal{S}}(i,i,1) = \frac{1}{I_3}\sum_{j=1}^{I_3} \widetilde{\boldsymbol{\mathcal{S}}}(i,i,j), \tag{S1.3}$$

and the $\widetilde{\boldsymbol{\mathcal{S}}}(:,:,j)$ is the singular value matrix of $\widetilde{\boldsymbol{\mathcal{A}}}(:,:,j)$. Therefore, the elements of $\boldsymbol{\mathcal{S}}(:,:,1)$ are regarded as the singular values of $\boldsymbol{\mathcal{A}}$.

Definition S1.7 *(Tensor tubal-rank, Lu et al. (2018)): For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times I_3}$, its*

tubal-rank is the number of nonzero singular value, that is

$$\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) := \#\{i, \boldsymbol{\mathcal{S}}(i,i,1) \neq 0\},$$

where "#" denotes the cardinality of a set.

It is easy to obtain $\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) \leq \mathrm{rank}(\overline{\boldsymbol{A}}) \leq I_3 \cdot \mathrm{rank_t}(\boldsymbol{\mathcal{A}})$; see Lu et al. (2018) for details.

Definition S1.8 *Define* $\|\boldsymbol{\mathcal{A}}\| := \|\overline{\boldsymbol{A}}\|_{op}$ *as the tensor spectral norm for* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$

Definition S1.9 *(t-TNN, Lu et al. (2018)): Let* $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}}$ *be the t-SVD of* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$*, and the tubal nuclear norm is defined as*

$$\|\boldsymbol{\mathcal{A}}\|_* := \sum_{i=1}^{r} \boldsymbol{\mathcal{S}}(i,i,1),$$

where $r = \mathrm{rank_t}(\boldsymbol{\mathcal{A}})$.

Note that the tensor spectral norm is the dual norm of t-TNN. It is easy to see that the following properties hold,

$$\|\boldsymbol{\mathcal{A}}\| = \|\overline{\boldsymbol{A}}\|_{op} = \|\widetilde{\boldsymbol{A}}\|_{op}, \|\boldsymbol{\mathcal{A}}\|_* = \frac{1}{I_3}\|\overline{\boldsymbol{A}}\|_* = \frac{1}{I_3}\|\widetilde{\boldsymbol{A}}\|_*. \tag{S1.4}$$

Proposition S1.1 *(Lu et al., 2018) Let* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ *with* $\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) = r$*,*

and its skinny t-SVD be $\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}}$ *where* $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{I_1 \times r \times I_3}$, $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r \times r \times I_3}$ *and* $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{I_2 \times r \times I_3}$. *The subdifferential of t-TNN is* $\partial \|\boldsymbol{\mathcal{A}}\|_* = \{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}} + \boldsymbol{\mathcal{W}} : \boldsymbol{\mathcal{U}}^{\mathsf{H}} * \boldsymbol{\mathcal{W}} = 0, \boldsymbol{\mathcal{W}} * \boldsymbol{\mathcal{V}} = 0, \|\boldsymbol{\mathcal{W}}\| \leq 1\}$.

With the above discussion, given $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$,

$$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle \leq \frac{1}{I_3} \langle \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}} \rangle \leq \frac{1}{I_3} \|\widetilde{\boldsymbol{A}}\|_* \|\widetilde{\boldsymbol{B}}\| = \|\boldsymbol{\mathcal{A}}\|_* \|\boldsymbol{\mathcal{B}}\|.$$

Moreover, let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with $\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) = r$,

$$\|\boldsymbol{\mathcal{A}}\|_* = \frac{1}{I_3} \|\widetilde{\boldsymbol{A}}\|_* \leq \frac{1}{I_3} \sqrt{r'} \|\widetilde{\boldsymbol{A}}\|_F \leq \sqrt{r} \|\boldsymbol{\mathcal{A}}\|_F,$$

where r' is the rank of $\widetilde{\boldsymbol{A}}$ and $r' \leq I_3 r$ holds. To prove that t-TNN is decomposable, the additivity of nuclear norm is needed.

Proposition S1.2 *(Recht et al., 2010) For any* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *with the same dimension, if* $\boldsymbol{A}\boldsymbol{B}^{\mathsf{H}} = \boldsymbol{A}^{\mathsf{H}}\boldsymbol{B} = 0$, *then* $\|\boldsymbol{A} + \boldsymbol{B}\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$.

Lemma S1.1 *For any* $\boldsymbol{\mathcal{A}}$ *and* $\boldsymbol{\mathcal{B}}$ *with the same dimension, if* $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^{\mathsf{H}} = \boldsymbol{\mathcal{A}}^{\mathsf{H}} * \boldsymbol{\mathcal{B}} = 0$, *we have* $\|\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}\|_* = \|\boldsymbol{\mathcal{A}}\|_* + \|\boldsymbol{\mathcal{B}}\|_*$.

Proof of Lemma S1.1: *Note that,*

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^{\mathsf{H}} = \boldsymbol{\mathcal{A}}^{\mathsf{H}} * \boldsymbol{\mathcal{B}} = 0 \rightarrow \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{B}}^{\mathsf{H}} = \widetilde{\boldsymbol{A}}^{\mathsf{H}}\widetilde{\boldsymbol{B}} = 0.$$

Thus,

$$\|\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\|_*=\frac{1}{I_3}\|(\widetilde{\boldsymbol{A}+\boldsymbol{B}})\|_*=\frac{1}{I_3}(\|\widetilde{\boldsymbol{A}}\|_*+\|\widetilde{\boldsymbol{B}}\|_*)=\|\boldsymbol{\mathcal{A}}\|_*+\|\boldsymbol{\mathcal{B}}\|_*.$$

■

Theorem S1.1 *For any $\boldsymbol{\mathcal{A}}\in\mathscr{U}^\perp$ and $\boldsymbol{\mathcal{B}}\in\mathscr{B}$, of the same dimension, with tubal rank r, we have $\|\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\|_*=\|\boldsymbol{\mathcal{A}}\|_*+\|\boldsymbol{\mathcal{B}}\|_*$. where*

$$\mathscr{U}=\left\{\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{M}}+\boldsymbol{\mathcal{N}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}:\boldsymbol{\mathcal{M}}\in\mathbb{R}^{r\times I_2\times I_3},\boldsymbol{\mathcal{N}}\in\mathbb{R}^{I_1\times r\times I_3}\right\},$$
$$\mathscr{B}=\left\{\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{U}}^{\mathsf{H}}*\boldsymbol{\mathcal{Z}}+\boldsymbol{\mathcal{Z}}*\boldsymbol{\mathcal{V}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}-\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{U}}^{\mathsf{H}}*\boldsymbol{\mathcal{Z}}*\boldsymbol{\mathcal{V}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}:\boldsymbol{\mathcal{Z}}\in\mathbb{R}^{I_1\times I_2\times I_3}\right\}.$$

Proof of Theorem S1.1 *Note that, $\boldsymbol{\mathcal{B}}\in\mathbb{R}^{I_1\times I_2\times I_3}$ has the skinny t-SVD, that is $\boldsymbol{\mathcal{B}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{S}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}$ where $\boldsymbol{\mathcal{U}}\in\mathbb{R}^{I_1\times r\times I_3}$, $\boldsymbol{\mathcal{S}}\in\mathbb{R}^{r\times r\times I_3}$ and $\boldsymbol{\mathcal{V}}\in\mathbb{R}^{I_2\times r\times I_3}$. Define $\mathscr{U}=\left\{\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{M}}+\boldsymbol{\mathcal{N}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}:\boldsymbol{\mathcal{M}}\in\mathbb{R}^{r\times I_2\times I_3},\boldsymbol{\mathcal{N}}\in\mathbb{R}^{I_1\times r\times I_3}\right\}$ and $\mathscr{B}=\left\{\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{U}}^{\mathsf{H}}*\boldsymbol{\mathcal{Z}}+\boldsymbol{\mathcal{Z}}*\boldsymbol{\mathcal{V}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}-\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{U}}^{\mathsf{H}}*\boldsymbol{\mathcal{Z}}*\boldsymbol{\mathcal{V}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}:\boldsymbol{\mathcal{Z}}\in\mathbb{R}^{I_1\times I_2\times I_3}\right\}$, thus, $\mathscr{B}^\perp=\left\{(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{U}}^{\mathsf{H}})*\boldsymbol{\mathcal{Z}}*(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{V}}*\boldsymbol{\mathcal{V}}^{\mathsf{H}}):\boldsymbol{\mathcal{Z}}\in\mathbb{R}^{I_1\times I_2\times I_3}\right\}$. It is easy to see that $\mathscr{B}\subseteq\mathscr{U}$ and $\mathscr{U}^\perp\subseteq\mathscr{B}^\perp$. Therefore, $\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{B}}^{\mathsf{H}}=0$ and $\boldsymbol{\mathcal{A}}^{\mathsf{H}}*\boldsymbol{\mathcal{B}}=0$. Then, by Lemma S1.1, we have $\|\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\|_*=\|\boldsymbol{\mathcal{A}}\|_*+\|\boldsymbol{\mathcal{B}}\|_*$.* ■

S2 Proofs and Lemmas for results in Section 2

The theoretical developments are following the roadmap below: The Lemma S2.1 approximates the difference between empirical risk and population risk via Rademacher complexity. The Lemma S2.2 illustrates that the statistical error $\|\sum_i L_h'(u_i^*)y_i\boldsymbol{\mathcal{X}}_i/n\|$ is bounded by λ through a kind of Talagrand's concentration inequality (Theorem 4.2 of Bartlett et al. (2005)). Furthermore, Theorem 1 investigates the upper bound of estimation error with the aid of Lemma S2.1 and S2.2. Then we study the size of $\mathbb{E}[\|\sum_i \sigma_i\boldsymbol{\mathcal{X}}_i\|]$ and $\mathbb{E}[\|\sum_i L_h'(u_i^*)y_i\boldsymbol{\mathcal{X}}_i\|]$ in **Assumption** 3 through the Lemma S2.3, and the tail probability $P(\|\boldsymbol{\mathcal{X}}\| > M)$ with a proper constant M via the Lemma S2.4, respectively. With the above results, Corollary 1 presents the statistical rate of the RSSTM estimator. Next, we show that the RSSTM estimator enjoys the low rank guarantee of the order $O(r)$. Finally, we investigate the convergence analysis of the proposed algorithm.

To simplify notations, the empirical measure on the n observations is denoted by P_n and the corresponding population counterpart is indicated by P. That is, with any function f on the same probability space as $(\boldsymbol{\mathcal{X}}, y)$, we have $Pf = \int f dP = \mathbb{E}f$.

Lemma S2.1 *Suppose **Assumption** 1-4 hold, for* $\forall M > 0$ *with* $\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|^2]P(\|\boldsymbol{\mathcal{X}}\| >$

$M) \leq \lambda/4$, *we have*

$$\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}} \left|(P_n-P)[L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))-L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))]\right| \leq C\lambda t\sqrt{r},$$

with probability $1 - Ce^{-Cn\lambda/M} - CnP(\|\boldsymbol{\mathcal{X}}\| > M)$, *where* $\mathscr{T} := \{(\delta, \boldsymbol{\mathcal{D}}) : \|\boldsymbol{\mathcal{D}}^{\perp}\|_* \leq 3\|\boldsymbol{\mathcal{D}}^0\|_* + |\delta|, \mathbb{E}[(\delta + \langle\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{D}}\rangle)^2] \leq t^2\}$.

Proof of Lemma S2.1: *For simplicity of notations, let* $f(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}, y) = L_h(y(\beta^* + \delta + \langle\boldsymbol{\mathcal{B}}^* + \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}\rangle)) - L_h(y(\beta^* + \langle\boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}\rangle))$, *and define* $\mathcal{F} = \{f(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}, y) : (\delta, \boldsymbol{\mathcal{D}}) \in \mathscr{T}\}$. *It is trivial that* $\mathrm{Var}(f(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}, y)) \leq C\mathrm{Var}(|y(\delta + \langle\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{D}}\rangle)|) \leq Ct^2$ *holds for* $f \in \mathcal{F}$ *by virtue of the Lipschitz continuity of the smoothed function* $L_h(\cdot)$. *Then the Markov inequality suggests that,*

$$P(|(P_n - P)f| < a) \geq 1 - \frac{Ct^2}{na^2},\ \forall\ a > 0.$$

The following inequality holds via the symmetrization Lemma (Pollard, 1984, Sec. II.3.8),

$$P(\sup_{f\in\mathcal{F}} |(P_n - P)f| > 2a) \leq \frac{P\big(\sup_{f\in\mathcal{F}} |(P_n - P_n')f| > a\big)}{1 - Ct^2/(na^2)},$$

where P_n' *denotes the empirical measure on the independent copies* $\{(y_i', \boldsymbol{\mathcal{X}}_i')\}_{i=1}^n$ *of* $\{(y_i, \boldsymbol{\mathcal{X}}_i)\}_{i=1}^n$.

Let $f_M = f\mathrm{I}\{\|\boldsymbol{\mathcal{X}}\| \leq M\}$ *for* $f \in \mathcal{F}$ *where* $\mathrm{I}\{\cdot\}$ *denotes the indicator*

function. Note that $\sup_{f\in\mathcal{F}}(P_n - P_n')f_M$ *and* $\sup_{f\in\mathcal{F}}\sum_i \sigma_i f_M(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i, y_i)/n$ *are identically distributed. By* $\|\boldsymbol{\mathcal{D}}\|_* \le C(\|\boldsymbol{\mathcal{D}}^0\|_* + |\delta|) \le Ct\sqrt{r}$ *due to the low tubal rank, we have* $|f_M| \le CMt\sqrt{r}$*. Then, using* $\mathrm{Var}(f) \le Ct^2$ *and the Talagrand's concentration inequality (see Theorem A.2 of Bartlett et al. (2005)), we have*

$$P\Big(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i \sigma_i f_M(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i, y_i) > \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i \sigma_i f_M(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i, y_i)\Big] + \frac{CMx}{n}t\sqrt{r}\Big) \le e^{-x}. \tag{S2.5}$$

On the other hand, we also have

$$\begin{aligned}
&\mathbb{E}\big(\|\sum_i \sigma_i y_i \boldsymbol{\mathcal{X}}_i \mathrm{I}\{\|\boldsymbol{\mathcal{X}}_i\| > M\}\|\big) = n\mathbb{E}\big(\sup_{\|\boldsymbol{\mathcal{A}}\|\le 1}\sigma y\langle \boldsymbol{\mathcal{X}}\mathrm{I}\{\|\boldsymbol{\mathcal{X}}\| > M\}, \boldsymbol{\mathcal{A}}\rangle\big)\\
&= n\mathbb{E}\big(\sup_{\|\boldsymbol{\mathcal{A}}\|\le 1}\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}\rangle \mathrm{I}\{\|\boldsymbol{\mathcal{X}}\| > M\}\big) \le n\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|^2]P(\|\boldsymbol{\mathcal{X}}\| > M) \le Cn\lambda,
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i \sigma_i f_M(\delta, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i, y_i)\Big] \le \mathbb{E}\Big[\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}\frac{1}{n}\sum_i \sigma_i y_i\big(\delta + \langle \boldsymbol{\mathcal{X}}_i \mathrm{I}\{\|\boldsymbol{\mathcal{X}}_i\| \le M\}, \boldsymbol{\mathcal{D}}\rangle\big)\Big]\\
&\le \mathbb{E}\Big[\frac{1}{n}\|\sum_i \sigma_i y_i \boldsymbol{\mathcal{X}}_i \mathrm{I}\{\|\boldsymbol{\mathcal{X}}_i\| \le M\}\|\Big]\cdot\Big(\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}\|\boldsymbol{\mathcal{D}}\|_*\Big) + \mathbb{E}\Big[\sup_{\delta}\frac{1}{n}\sum_i \sigma_i y_i \delta\Big]\\
&\le \mathbb{E}\Big[\frac{1}{n}\|\sum_i \sigma_i y_i \boldsymbol{\mathcal{X}}_i\|\Big]\cdot\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}\|\boldsymbol{\mathcal{D}}\|_* + \mathbb{E}\Big[\frac{1}{n}\|\sum_i \sigma_i y_i \boldsymbol{\mathcal{X}}_i \mathrm{I}\{\|\boldsymbol{\mathcal{X}}_i\| > M\}\|\Big]\cdot\Big(\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}\|\boldsymbol{\mathcal{D}}\|_*\Big)\\
&+ \mathbb{E}\Big[\sup_{\delta}\frac{1}{n}\sum_i \sigma_i y_i \delta\Big] \le C\lambda t\sqrt{r}.
\end{aligned} \tag{S2.6}$$

Choosing $x \asymp n\lambda/M$ *and* $a = C\lambda t\sqrt{r}$*, with* (S2.5) *and* (S2.6)*, we have*

$$P(\sup_{f\in\mathcal{F}}(P_n - P_n')f_M > a/3) \le e^{-x}.$$

Then

$$P(\sup_{f\in\mathcal{F}}P_n(f - f_M) > a/3) = P(\sup_{f\in\mathcal{F}}P_n'(f - f_M) > a/3)$$
$$\le P(\max_i\|\boldsymbol{\mathcal{X}}_i\| > M) \le nP(\|\boldsymbol{\mathcal{X}}\| > M).$$

This implies $P(\sup_{f\in\mathcal{F}}(P_n - P)f > a) \le Ce^{-x} + CnP(\|\boldsymbol{\mathcal{X}}\| > M)$. ■

Lemma S2.2 *For any* $M > 0$ *satisfying* $\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|^2]P(\|\boldsymbol{\mathcal{X}}\| > M) \le \lambda/4$*, with probability* $1 - e^{-Cn\lambda/M} - nP(\|\boldsymbol{\mathcal{X}}\| > M)$*,*

$$\lambda \ge 2\|\sum_i L_h'(u_i^*)y_i\boldsymbol{\mathcal{X}}_i/n\|.$$

Proof of Lemma S2.2: *For brevity, we indicate* $\nu_i = L_h'(u_i^*)y_i$*. Using the boundedness of* L_h' *and optimality condition, we have* $|\nu_i| \le 1$ *and* $\mathbb{E}[\nu_i\boldsymbol{\mathcal{X}}_i] = 0$ *. In fact,*

$$\|\sum_i \nu_i\boldsymbol{\mathcal{X}}_i/n\| = \sup_{\|\boldsymbol{\mathcal{A}}\|_*\le 1}\frac{1}{n}\sum_i \nu_i\langle\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}}\rangle.$$

Let $\mathscr{G} := \{g : g(\nu, \boldsymbol{\mathcal{X}}) = \nu\langle\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}\rangle, \|\boldsymbol{\mathcal{A}}\|_* \le 1\}$ *and* $g_M = g\mathrm{I}\{|g| \le M\} -$

$M\mathrm{I}\{g < -M\} + M\mathrm{I}\{g > M\}$. *Then, we have*

$$\|\sum_i \nu_i \boldsymbol{\mathcal{X}}_i/n\| = \sup_{g\in\mathscr{G}}(P_n - P)g \le \sup_{g\in\mathscr{G}}(P_n - P)g_M + |\sup_{g\in\mathscr{G}}(P_n - P)(g - g_M)|.$$

For the first term of the above equation, with probability $1 - e^{-x}$*,*

$$\sup_{g\in\mathscr{G}}(P_n - P)g_M \le C\mathbb{E}[\sup_{g\in\mathscr{G}}(P_n - P)g_M] + \frac{CMx}{n} \le C\mathbb{E}[\|\sum_i \nu_i \boldsymbol{\mathcal{X}}_i/n\|] + \frac{CMx}{n},$$

where the first inequality used the improved version of Talagrand's concentration inequality (see Theorem A.2 of Bartlett et al. (2005)) and the last one used the contraction property of Rademacher complexity for Lipschitz functions (see Theorem 2.2 of Koltchinskii (2011)). Taking $x \asymp n\lambda/M$*, we have* $\sup_{g\in\mathscr{G}}(P_n - P)g_M \le \lambda/4$.

Thus

$$P\big(\|\sum_i \nu_i \boldsymbol{\mathcal{X}}_i/n\| \le \lambda/4 + |\sup_{g\in\mathscr{G}}(P_n - P)(g - g_M)|\big) \ge 1 - e^{-Cn\lambda/M}.$$

On the other hand, with probability $1 - nP(\|\boldsymbol{\mathcal{X}}\| > M)$*,*

$$|\sup_{g\in\mathscr{G}}(P_n - P)(g - g_M)| \le \sup_{g\in\mathscr{G}}|P(g - g_M)| \le \mathbb{E}[\|\boldsymbol{\mathcal{X}}\|^2]P(\|\boldsymbol{\mathcal{X}}\| > M) \le \lambda/4.$$

Then,

$$P(\|\sum_i L_h'(u_i^*)y_i\boldsymbol{\mathcal{X}}_i/n\| \le \lambda/2) \ge 1 - e^{-Cn\lambda/M} - nP(\|\boldsymbol{\mathcal{X}}\| > M).$$

■

Proof of Theorem 1: *By definition of the RSSTM estimator in Section 2.2, we have*

$$\sum_{i=1}^n L_h(y_i(\widehat{\beta} + \langle\widehat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{X}}_i\rangle)) + n\lambda\|\widehat{\boldsymbol{\mathcal{B}}}\|_* \le \sum_{i=1}^n L_h(y_i(\beta^* + \langle\boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}_i\rangle)) + n\lambda\|\boldsymbol{\mathcal{B}}^*\|_*. \tag{S2.7}$$

*On the other hand, using the convexity of smoothed loss function L_h and the inequality $\langle\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\rangle \le \|\boldsymbol{\mathcal{A}}\|\|\boldsymbol{\mathcal{B}}\|_*$, **Assumption** 3 and the result of Lemma S2.2, we have*

$$\sum_{i=1}^n \{L_h(y_i(\widehat{\beta} + \langle\widehat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{X}}_i\rangle)) - L_h(y_i(\beta^* + \langle\boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}_i\rangle))\} \ge \sum_{i=1}^n L_h'(u_i^*)y_i(\langle\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}_i\rangle + \widehat{\beta} - \beta^*)$$
$$\ge -\|\sum_{i=1}^n L_h'(u_i^*)y_i\boldsymbol{\mathcal{X}}_i\|\|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\|_* - |\sum_{i=1}^n L_h'(u_i^*)y_i||\widehat{\beta} - \beta| \ge -\frac{n\lambda}{2}(\|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\|_* + |\widehat{\beta} - \beta^*|),$$

where we use that $n\lambda \ge C\sqrt{n\log n} \ge 2|\sum_{i=1}^n L_h'(u_i^*)y_i|$ with high probability by Hoeffding's inequality. Recall (S2.7)*,*

$$\sum_{i=1}^n L_h(y_i(\widehat{\beta} + \langle\widehat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{X}}_i\rangle)) - \sum_{i=1}^n L_h(y_i(\beta^* + \langle\boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}_i\rangle)) \le n\lambda(\|\boldsymbol{\mathcal{B}}^*\|_* - \|\widehat{\boldsymbol{\mathcal{B}}}\|_*),$$

which yields that

$$-\frac{\lambda}{2}(\|\widehat{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}^*\|_*+|\widehat{\beta}-\beta^*|)\leq\lambda(\|\boldsymbol{\mathcal{B}}^*\|_*-\|\widehat{\boldsymbol{\mathcal{B}}}\|_*).$$

Let $\widehat{\boldsymbol{\mathcal{D}}}=\widehat{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}^*$ *and* $\widehat{\boldsymbol{\mathcal{D}}}^{\perp}$ *be the projection of* $\widehat{\boldsymbol{\mathcal{D}}}$ *on* $\mathscr{U}^{\perp}$*, and* $\widehat{\boldsymbol{\mathcal{D}}}^{0}=\widehat{\boldsymbol{\mathcal{D}}}-\widehat{\boldsymbol{\mathcal{D}}}^{\perp}$. *Also, denote* $\widehat{\delta}=\widehat{\beta}-\beta^*$. *Thus, by the decomposability of t-TNN, the above inequality implies*

$$\begin{aligned}&-\frac{1}{2}(\|\widehat{\boldsymbol{\mathcal{D}}}^{0}\|_*+\|\widehat{\boldsymbol{\mathcal{D}}}^{\perp}\|_*+|\widehat{\delta}|)\leq\|\boldsymbol{\mathcal{B}}^*\|_*-\|\boldsymbol{\mathcal{B}}^*+\widehat{\boldsymbol{\mathcal{D}}}^{\perp}\|_*+\|\widehat{\boldsymbol{\mathcal{D}}}^{0}\|_*\\&\leq\|\boldsymbol{\mathcal{B}}^*\|_*-\|\boldsymbol{\mathcal{B}}_1^*+\widehat{\boldsymbol{\mathcal{D}}}^{\perp}\|_*+\|\widehat{\boldsymbol{\mathcal{D}}}^{0}\|_*=-\|\widehat{\boldsymbol{\mathcal{D}}}^{\perp}\|_*+\|\widehat{\boldsymbol{\mathcal{D}}}^{0}\|_*,\end{aligned}$$

which yields

$$\|\widehat{\boldsymbol{\mathcal{D}}}^{\perp}\|_*\leq3\|\widehat{\boldsymbol{\mathcal{D}}}^{0}\|_*+|\widehat{\delta}|.$$

That is, $(\widehat{\delta},\widehat{\boldsymbol{\mathcal{D}}})\in\mathscr{H}:=\{(\delta,\boldsymbol{\mathcal{D}}):\|\boldsymbol{\mathcal{D}}^{\perp}\|_*\leq3\|\boldsymbol{\mathcal{D}}^{0}\|_*+|\delta|\}$.*Then, by Lemma S2.1, we have, with probability* $1-e^{-Cn\lambda/M}-nP(\|\boldsymbol{\mathcal{X}}\|>M)$,

$$\begin{aligned}\inf_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}\Big\{&\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))-\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))\\&-\mathbb{E}[L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))]+\mathbb{E}[L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))]\Big\}\\&\geq-C\lambda t\sqrt{r}.\end{aligned}\tag{S2.8}$$

Suppose $\mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]\geq t^2$ *for some* $t>0$*, by* (S2.7)*, then we have*

$$\inf_{\substack{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{H}\\ \mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]\geq t^2}}\Big\{\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))-\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))$$
$$+\lambda\|\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}}\|_*-\lambda\|\boldsymbol{\mathcal{B}}^*\|_*\Big\}<0,$$

and by the convexity of the objective function,

$$\inf_{\substack{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{H}\\ \mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]=t^2}}\Big\{\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))-\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))$$
$$+\lambda\|\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}}\|_*-\lambda\|\boldsymbol{\mathcal{B}}^*\|_*\Big\}<0. \tag{S2.9}$$

Thus, for some $(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{H}$ *satisfying* $\mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]=t^2$*,*

$$\begin{aligned}
&\mathbb{E}[L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))]-\mathbb{E}[L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))]\\
&\leq\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\delta+\langle\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}\rangle))-\frac{1}{n}\sum_{i=1}^{n}L_h(y(\beta^*+\langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}\rangle))+C\lambda t\sqrt{r}\\
&\leq\lambda(\|\boldsymbol{\mathcal{B}}^*\|_*-\|\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}}^{\perp}\|_*+\|\boldsymbol{\mathcal{D}}^0\|_*)+C\lambda t\sqrt{r}\\
&\leq C\lambda(\|\boldsymbol{\mathcal{D}}^0\|_*+t\sqrt{r})\leq C\lambda(\sqrt{r}\|\boldsymbol{\mathcal{D}}\|_F+t\sqrt{r})\leq C\lambda t\sqrt{r},
\end{aligned}$$

where the first two inequalities are based on (S2.8) *and* (S2.9)*, the 3rd inequality uses the decomposibility of t-TNN, the penultimate inequality uses the low rank, and the last one applies* $\|\boldsymbol{\mathcal{D}}\|_F\leq Ct$ *by* $\mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]=t^2$ *and* ***Assumption*** *2, which suggests the first result. Then, under* ***Assumption*** *2, we can immediately*

get the second result. By $\|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\|_* \leq \|(\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*)^{\perp}\|_* + \|(\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*)^0\|_* \leq C(\|(\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*)^0\|_* + |\widehat{\beta} - \beta^*|) \leq C\sqrt{r}(\|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\|_F + |\widehat{\beta} - \beta^*|)$ *due to* $\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^* \in \mathscr{H}$*, we have the last result.* ■

Lemma S2.3 *Under* ***Assumption*** *2, we have*

$$\mathbb{E}[\|\sum_i L_h'(u_i^*) y_i \boldsymbol{\mathcal{X}}_i\|] \leq C\sqrt{n(I_1 I_3 \vee I_2 I_3)} \text{ and } \mathbb{E}[\|\sum_i \sigma_i \boldsymbol{\mathcal{X}}_i\|] \leq C\sqrt{n(I_1 I_3 \vee I_2 I_3)}.$$

Proof of Lemma S2.3: *Similar to Lemma S2.2, let* $\nu_i = L_h'(u_i^*) y_i$ *which is bounded. From Appendix S1, it can be seen that* $\|\sum_i \nu_i \boldsymbol{\mathcal{X}}_i\| = \|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op}$ *where* $\overline{\boldsymbol{X}}_i \in \mathbb{R}^{I_1 I_3 \times I_2 I_3}$*. It is trivial that* $\overline{\boldsymbol{X}}_i, i = 1, \cdots, n$*, are sub-Gaussian matrices via the sub-Gaussianity of* $\boldsymbol{x}_i = \text{vec}(\boldsymbol{\mathcal{X}}_i)$*. With the boundedness of* ν_i*, let* $\bar{\boldsymbol{x}}_i = \text{vec}(\overline{\boldsymbol{X}}_i)$ *and* $\nu_i \bar{\boldsymbol{x}}_i$ *is also sub-Gaussian. With the covering of unit spheres* $\boldsymbol{u}_i$ *and* $\boldsymbol{v}_j$ *for* $\sum_i \nu_i \overline{\boldsymbol{X}}_i$*, defined in the proof of Lemma S2.4, we have, for any* $t > 0$*,*

$$P(\|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} > 2a) \leq \sum_{i,j} P\big(\boldsymbol{u}_i^{\mathsf{T}} (\sum_i \nu_i \overline{\boldsymbol{X}}_i) \boldsymbol{v}_j > a\big) \leq \sum_{i,j} P\big((\boldsymbol{v}_j \otimes \boldsymbol{u}_i)^{\mathsf{T}} \text{vec}(\sum_i \nu_i \overline{\boldsymbol{X}}_i)\big)$$

$$\leq \sum_{i,j} e^{-ta} \mathbb{E}[e^{t(\boldsymbol{v}_j \otimes \boldsymbol{u}_i)^{\mathsf{T}} \text{vec}(\sum_i \nu_i \overline{\boldsymbol{X}}_i)}] \leq 20^{I_1 I_3 + I_2 I_3} e^{-ta + Cnt^2} \leq 20^{I_1 I_3 + I_2 I_3} e^{-ta + Cnt^2}.$$

Taking $t \asymp a/n$ *and* $a = C\sqrt{n((I_1 I_3 \vee I_2 I_3) - \log\tau)}$ *for any* $\tau > 0$*, we have*

$$P(\|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} > C\sqrt{n((I_1 I_3 \vee I_2 I_3) - \log\tau)}) \leq \tau.$$

Denote $t = \log(1/\tau)$ *and* $a = C\sqrt{n(I_1I_3 \vee I_2I_3)}$*, which implies* $\|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} > at^{1/2}$ *with probability at least* e^{-t}*. Thus, we have*

$$\mathbb{E}[\|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op}] \le \sum_{t=1}^{\infty} \mathbb{E}\big[\|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} \cdot \mathrm{I}\{a(t-1)^{1/2} < \|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} \le at^{1/2}\}\big]$$
$$\le \sum_{t=1}^{\infty} at^{1/2} P\big(a(t-1)^{1/2} < \|\sum_i \nu_i \overline{\boldsymbol{X}}_i\|_{op} \le at^{1/2}\big) \le \sum_{t=1}^{\infty} at^{1/2} e^{-t+1} = C\sqrt{n(I_1I_3 \vee I_2I_3)}.$$

We can get $\mathbb{E}[\|\sum_i \sigma_i \boldsymbol{\mathcal{X}}_i\|] \le C\sqrt{n(I_1I_3 \vee I_2I_3)}$ *in the same way.* ■

Lemma S2.4 *Under* ***Assumption*** *2, we have* $P\big(\|\boldsymbol{\mathcal{X}}\| > C\sqrt{(I_1I_3 \vee I_2I_3)\log n}\big) \le Ce^{-C(I_1I_3 \vee I_2I_3)\log n}$.

Proof of Lemma S2.4: *We have* $\|\boldsymbol{\mathcal{X}}\| = \|\overline{\boldsymbol{X}}\|_{op}$ *where* $\overline{\boldsymbol{X}} \in \mathbb{R}^{I_1I_3 \times I_2I_3}$*. Let* $\{\boldsymbol{u}_i\}_{i=1}^{M_1}$ *and* $\{\boldsymbol{v}_j\}_{i=1}^{M_2}$ *be the* $1/4$ *covering of the unit sphere in* $\mathbb{R}^{I_1I_3}$ *and* $\mathbb{R}^{I_2I_3}$*, for* $\overline{\boldsymbol{X}}$*, respectively, with* $M_1 \le 20^{I_1I_3}$ *and* $M_2 \le 20^{I_2I_3}$ *referred from Lemma 2.5 in Geer (2000). Thus, for any* $\boldsymbol{u}, \boldsymbol{v}$ *if* $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$*, there exists* $\boldsymbol{u}_i, \boldsymbol{v}_j$ *in the covering satisfy* $\|\boldsymbol{u} - \boldsymbol{u}_i\| \le 1/4, \|\boldsymbol{v} - \boldsymbol{v}_j\| \le 1/4$*, and then*

$$\boldsymbol{u}^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v} = \boldsymbol{u}^{\mathsf{T}}\overline{\boldsymbol{X}}(\boldsymbol{v} - \boldsymbol{v}_j) + (\boldsymbol{u} - \boldsymbol{u}_i)^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j \le \frac{1}{2}\|\overline{\boldsymbol{X}}\|_{op} + \boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j.$$

Then, we have

$$\|\overline{\boldsymbol{X}}\|_{op} = \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} \boldsymbol{u}^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v} \le \frac{1}{2}\|\overline{\boldsymbol{X}}\|_{op} + \max_{\boldsymbol{u}_i, \boldsymbol{v}_j} \boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j,$$

which yields $\|\overline{\boldsymbol{X}}\|_{op} \leq 2\max_{\boldsymbol{u}_i,\boldsymbol{v}_j} \boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j$. *Thus, for any* $t > 0$, *due to the sub-Gaussianity of* $\boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j = (\boldsymbol{v}_j \otimes \boldsymbol{u}_i)^{\mathsf{T}}\mathrm{vec}(\overline{\boldsymbol{X}})$, *we have*

$$\mathbb{E}[e^{t\|\overline{\boldsymbol{X}}\|_{op}}] \leq \mathbb{E}\big[e^{2t\max_{\boldsymbol{u}_i,\boldsymbol{v}_j}\boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j}\big] \leq \mathbb{E}\big[\sum_{i,j} e^{2t\boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j}\big]$$

$$\leq 20^{I_1I_3+I_2I_3}\max_{\boldsymbol{u}_i,\boldsymbol{v}_j}\mathbb{E}[e^{2t\boldsymbol{u}_i^{\mathsf{T}}\overline{\boldsymbol{X}}\boldsymbol{v}_j}] \leq 20^{I_1I_3+I_2I_3}e^{Ct^2}.$$

By Markov's inequality,

$$P(\|\overline{\boldsymbol{X}}\|_{op} > s) \leq e^{-ts}20^{I_1I_3+I_2I_3}e^{Ct^2} \leq Ce^{-Cs^2}.$$

Taking $s = C\sqrt{(I_1I_3 \vee I_2I_3)\log n}$ *completes the proof.* ■

Proof of Theorem 2: *By the optimality of* $\widehat{\boldsymbol{\mathcal{B}}}$, *there exists* $\boldsymbol{\mathcal{Z}} \in \partial\|\widehat{\boldsymbol{\mathcal{B}}}\|_*$ *such that,* $\frac{1}{n}\sum_{i=1}^{n} L_h'(y_i(\hat{\beta} + \langle\widehat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{X}}_i\rangle))y_i\boldsymbol{\mathcal{X}}_i + \lambda\boldsymbol{\mathcal{Z}} = 0$. *Thus, we have*

$$\frac{1}{n}\sum_{i=1}^{n} L_h'(\hat{v}_i)y_i\overline{\boldsymbol{X}_i} + \lambda\overline{\boldsymbol{Z}} = 0, \tag{S2.10}$$

where $\hat{v}_i = y_i(\hat{\beta} + \frac{1}{I_3}\langle\overline{\widehat{\boldsymbol{B}}}, \overline{\boldsymbol{X}_i}\rangle)$ *and* $\overline{\boldsymbol{Z}} \in \partial\|\overline{\widehat{\boldsymbol{B}}}\|_*$. *Let* $\mathrm{rank_t}(\widehat{\boldsymbol{\mathcal{B}}}) = \hat{r}$, $\mathrm{rank}(\overline{\widehat{\boldsymbol{B}}}) = \hat{r}' \leq I_3\hat{r}$ *and its SVD given by* $\overline{\widehat{\boldsymbol{B}}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$, $\boldsymbol{U} \in \mathbb{R}^{I_1I_3\times\hat{r}'}$, $\boldsymbol{V} \in \mathbb{R}^{I_2I_3\times\hat{r}'}$. *Using (2.1) in Koltchinskii et al. (2011), it leads to*

$$\partial\|\overline{\widehat{\boldsymbol{B}}}\|_* = \{\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}} + \boldsymbol{U}^{\perp}\boldsymbol{W}(\boldsymbol{V}^{\perp})^{\mathsf{T}} : \|\boldsymbol{W}\|_{op} \leq 1\},$$

where $\boldsymbol{U}^{\perp} \in \mathbb{R}^{I_1 I_3 \times (I_1 I_3 - \hat{r}')}$ *and* $\boldsymbol{V}^{\perp} \in \mathbb{R}^{I_2 I_3 \times (I_2 I_3 - \hat{r}')}$ *are orthogonal matrices whose columns are orthogonal to those in* $\boldsymbol{U}$ *and* $\boldsymbol{V}$*, respectively. Let* $\boldsymbol{u}_j, \boldsymbol{v}_j, j = 1, \cdots, \hat{r}'$ *be the columns of* $\boldsymbol{U}$ *and* $\boldsymbol{V}$*. Pre(post)-multiplying* $\boldsymbol{u}_j$ *and* $\boldsymbol{v}_j$ *to* (S2.10) *yields,*

$$\frac{1}{n}\sum_{i=1}^{n} L_h'(\hat{v}_i) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j = -\lambda, \ j = 1, \cdots, \hat{r}'. \tag{S2.11}$$

According to the optimality condition of $(\beta^*, \boldsymbol{\mathcal{B}}^*)$, $\mathbb{E}[L_h'(v_i^*) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j] = 0$*. We write*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} L_h'(\hat{v}_i) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j \\
&= (P_n - P)[(L_h'(\hat{v}_i) - L_h'(v_i^*)) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j] + \mathbb{E}[(L_h'(\hat{v}_i) - L_h'(v_i^*)) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j] \\
&+ (P_n - P)(L_h'(v_i^*) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j) + \mathbb{E}[L_h'(v_i^*) y_i \boldsymbol{u}_j^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j] \\
&:= g_{1j} + g_{2j} + g_{3j}.
\end{aligned}$$

Let $\mathbf{g}_k = (g_{kj}, \cdots, g_{k\hat{r}'})^{\mathsf{T}}$ *for* $k = 1, 2, 3$ *and* $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{\hat{r}'})^{\mathsf{T}} \in \mathbb{R}^{\hat{r}'}$ *be any unit vector and* $\boldsymbol{\gamma} = (\boldsymbol{v}_1 \otimes \boldsymbol{u}_1, \cdots, \boldsymbol{v}_{\hat{r}'} \otimes \boldsymbol{u}_{\hat{r}'})\boldsymbol{\alpha}$*, we have* $\|\boldsymbol{\gamma}\|_2^2 = \sum_{j=1}^{\hat{r}'} \alpha_j^2 \|\boldsymbol{v}_j \otimes \boldsymbol{u}_j\|_2^2 = 1$*. Denote* $\mathscr{D} = \{\overline{\boldsymbol{B}} \in \mathbb{R}^{I_1 I_3 \times I_2 I_3} : \|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}^*}\|_F \leq d_n := C\sqrt{\hat{r}'(I_1 I_3 \vee I_2 I_3)/n}\}$*. Due to the Lipschitz continuity of* L_h' *and the*

sub-Gaussian property of $\boldsymbol{\mathcal{X}}$, then we have

$$\begin{aligned}\|\mathbf{g}_2\|_2 = \sup_{\|\alpha\|_2=1} \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{g}_2 &\leq \sup_{\|\boldsymbol{\gamma}\|_2=1,\overline{\boldsymbol{B}}\in\mathscr{D}} C\mathbb{E}[|\langle\overline{\boldsymbol{X}}_i,\overline{\boldsymbol{B}}-\overline{\boldsymbol{B}^*}\rangle+(\beta-\beta^*)||\boldsymbol{\gamma}^{\mathsf{T}}\mathrm{vec}(\overline{\boldsymbol{X}}_i)|] \\ &\leq \sup_{\|\boldsymbol{\gamma}\|_2=1,\overline{\boldsymbol{B}}\in\mathscr{D}} C\sqrt{\mathbb{E}|\langle\overline{\boldsymbol{X}}_i,\overline{\boldsymbol{B}}-\overline{\boldsymbol{B}^*}\rangle+(\beta-\beta^*)|^2}\sqrt{\mathbb{E}|\boldsymbol{\gamma}^{\mathsf{T}}\mathrm{vec}(\overline{\boldsymbol{X}}_i)|^2} \\ &\leq \sup_{\overline{\boldsymbol{B}}\in\mathscr{D}} C\Big(\|\overline{\boldsymbol{B}}-\overline{\boldsymbol{B}^*}\|_F+|\beta-\beta^*|\Big)\leq Cd_n.\end{aligned}$$

By similar arguments as Lemma S2.5, with high probability,

$$\|\mathbf{g}_1\|_2 \leq Cd_n\sqrt{\frac{\hat{r}'(I_1I_3\vee I_2I_3)\log n}{n}}+Cd_n\frac{(\hat{r}')^{3/2}(I_1I_3\vee I_2I_3)^2(\log n)^2}{n}.$$

By similar arguments as Lemma S2.4,

$$\|\mathbf{g}_3\|_2 \leq C\sqrt{\frac{(I_1I_3\vee I_2I_3)\log n}{n}},$$

with high probability. Hence,

$$\begin{aligned}&\lambda\sqrt{\hat{r}'} \leq \|\mathbf{g}_1\|_2+\|\mathbf{g}_2\|_2+\|\mathbf{g}_3\|_2 \\ &\leq Cd_n\sqrt{\frac{\hat{r}'(I_1I_3\vee I_2I_3)\log n}{n}}+Cd_n\frac{(\hat{r}')^{3/2}(I_1I_3\vee I_2I_3)^2(\log n)^2}{n} \\ &+Cd_n+C\sqrt{\frac{(I_1I_3\vee I_2I_3)\log n}{n}},\end{aligned}$$

which implies that $\hat{r}\leq Cr$ by taking $\lambda=C\sqrt{(I_1I_3\vee I_2I_3)/n}$. ■

Lemma S2.5 *Assume $\mathbb{E}(\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle+\delta)^4\leq C(\|\boldsymbol{\mathcal{D}}\|_F^4+|\delta|^4)$ for any $\boldsymbol{\mathcal{D}}\in\mathbb{R}^{I_1\times I_2\times I_3}$, $\delta\in$*

$\mathbb{R}$ *and some generic positive constant* C*, uniformly over* $\mathscr{A} = \{(\beta, \boldsymbol{\mathcal{B}}) : |\beta - \beta^*| + \|\boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{B}}^*\|_F \le d_n/\sqrt{I_3}, \mathrm{rank_t}(\boldsymbol{\mathcal{B}}) \le Cr\}$,

$$\sup_{(\beta,\boldsymbol{\mathcal{B}})\in\mathscr{A}} \left\| \frac{1}{n}\sum_{i=1}^{n}(L_h'(u_i) - L_h'(u_i^*))y_i\boldsymbol{\mathcal{X}}_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(L_h'(u_i) - L_h'(u_i^*))y_i\boldsymbol{\mathcal{X}}_i] \right\| \le Cd_n\Big(\sqrt{\frac{r(I_1I_3 \vee I_2I_3)\log n}{n}} + \frac{r^{3/2}(I_1I_3 \vee I_2I_3)^2(\log n)^2}{n}\Big).$$

with probability at least $1 - n^{-C}$.

Proof of Lemma S2.5: *We consider* $\mathscr{A}' = \{(\beta, \overline{\boldsymbol{B}}) : |\beta - \beta^*| + \|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^*\|_F \le d_n, \mathrm{rank_t}(\overline{\boldsymbol{B}}) \le Cr'\}$ *with* $r' \le I_3 r$. *It can be verified that* $(\beta, \overline{\boldsymbol{B}}) \in \mathscr{A}'$ *if* $(\beta, \boldsymbol{\mathcal{B}}) \in \mathscr{A}$. *Let* $q_i(\beta, \overline{\boldsymbol{B}}) = ((L_h'(v_i) - L_h'(v_i^*))y_i\overline{\boldsymbol{X}}_i - \mathbb{E}[(L_h'(v_i) - L_h'(v_i^*))y_i\overline{\boldsymbol{X}}_i]$. *Denote* $\mathcal{N}_{\mathscr{A}'}$ *as the* $d_n n^{-M}$*-covering of* $\mathscr{A}'$ *for a sufficiently large* $M > 0$, *then*

$$\begin{aligned}
&\sup_{(\beta,\overline{\boldsymbol{B}})\in\mathscr{A}'} \left\| \frac{1}{n}\sum_{i=1}^{n} q_i(\beta, \overline{\boldsymbol{B}}) \right\|_{op} \\
&\le \max_{(\beta',\overline{\boldsymbol{B}}')\in\mathcal{N}_{\mathscr{A}'}} \left\| \frac{1}{n}\sum_{i=1}^{n} q_i(\beta', \overline{\boldsymbol{B}}') \right\|_{op} \\
&+ \max_{(\beta',\overline{\boldsymbol{B}}')\in\mathcal{N}_{\mathscr{A}'}} \sup_{|\beta-\beta^*|+\|\overline{\boldsymbol{B}}-\overline{\boldsymbol{B}}^*\|_F \le d_n n^{-M}} \left\| \frac{1}{n}\sum_{i=1}^{n}\{q_i(\beta, \overline{\boldsymbol{B}}) - q_i(\beta', \overline{\boldsymbol{B}}')\} \right\|_{op} \\
&:= S_1 + S_2.
\end{aligned}$$

Next, we separately bound S_1 *and* S_2. *For any covering of unit spheres* $\{\boldsymbol{u}_i\}_{i=1}, \{\boldsymbol{v}_j\}_{j=1}$ *and the Lipschitz continuity, using Cauchy-Schwartz inequality and assumption*

$\mathbb{E}(\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{D}} \rangle + \delta)^4 \leq C(\|\boldsymbol{\mathcal{D}}\|_F^4 + |\delta|^4)$*, we have*

$$\mathbb{E}\big[(L_h'(v_i') - L_h'(v_i^*))y_i \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j\big]^2 \leq C\sqrt{\mathbb{E}[\langle \overline{\boldsymbol{X}_i}, \overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^* \rangle + |\beta' - \beta^*|]^4}\sqrt{\mathbb{E}[y_i \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j]^4}$$
$$\leq C(|\beta - \beta^*|^2 + \|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^*\|_F^2) \leq C d_n^2.$$

Moreover,

$$\big|(L_h'(v_i') - L_h'(v_i^*))y_i \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j\big| \leq C(\max_{1 \leq i \leq n} \|\overline{\boldsymbol{X}_i}\|_{op})\|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^*\|_*$$
$$\leq C\sqrt{r'} d_n \sqrt{(I_1 I_3 \vee I_2 I_3)\log n},$$

where the last result uses that $\max_{1 \leq i \leq n} \|\overline{\boldsymbol{X}_i}\|_{op} \leq C\sqrt{(I_1 I_2 \vee I_2 I_3)\log n}$ *with high probability by similar arguments as that in Lemma S2.4. Hence, by Bernstein's inequality, with probability at least* $1 - 2e^{-z}$ *for any* $z > 0$*, we have*

$$\Big|\frac{1}{n}\sum_{i=1}^{n}(L_h'(v_i') - L_h'(v_i^*))y_i \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j - \mathbb{E}\big[(L_h'(v_i') - L_h'(v_i^*))y_i \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j\big]\Big|$$
$$\leq C\sqrt{\frac{z d_n^2}{n}} + C\frac{z\sqrt{r'} d_n (I_1 I_3 \vee I_2 I_3)}{n} := \psi(z),$$

By Lemma 5.2 in Vershynin (2011), the covering number satisfies $|\mathcal{N}_{\mathscr{A}'}| \leq (1 +$

$2n^M)^{Cr'(I_1I_3+I_2I_3)}$. *Then we have*

$$\begin{aligned}
P(S_1 \geq \psi(z)) &= P\left(\max_{(\beta', \overline{\boldsymbol{B}}') \in \mathcal{N}_{\mathscr{A}'}} \left\| \frac{1}{n} \sum_{i=1}^{n} q_i(\beta', \overline{\boldsymbol{B}}') \right\|_{op} \geq \psi(z) \right) \\
&\leq Cn^{CMr'(I_1I_3 \vee I_2I_3)} \max_{(\beta', \overline{\boldsymbol{B}}') \in \mathcal{N}_{\mathscr{A}'}} P\left(\max_{\boldsymbol{u}_i, \boldsymbol{v}_j} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i^{\mathsf{T}} q_i(\beta', \overline{\boldsymbol{B}}') \boldsymbol{v}_j \geq \psi(z) \right) \\
&\leq Cn^{CMr'(I_1I_3 \vee I_2I_3)} 20^{I_1I_3 \vee I_2I_3} e^{-z}.
\end{aligned}$$

Taking $z = Cr'(I_1I_3 \vee I_2I_3)\log n$*, with probability at least* $1 - n^{-C}$*,*

$$S_1 \leq Cd_n \sqrt{\frac{r'(I_1I_3 \vee I_2I_3)\log n}{n}} + Cd_n \frac{(r')^{3/2}(I_1I_3 \vee I_2I_3)^2(\log n)^2}{n}. \quad \text{(S2.12)}$$

Next we bound S_2*. Denote*

$$w(\overline{\boldsymbol{X}_i}) = \sup_{|\beta - \beta^*| + \|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^*\|_F \leq d_n n^{-M}} [q_i(\beta, \overline{\boldsymbol{B}}) - q_i(\beta', \overline{\boldsymbol{B}}')].$$

Similarly, it leads to

$$\begin{aligned}
|\boldsymbol{u}_i^{\mathsf{T}} w(\overline{\boldsymbol{X}_i}) \boldsymbol{v}_j| &\leq \left| \sup_{|\beta - \beta^*| + \|\overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^*\|_F \leq d_n n^{-M}} \left| \langle \overline{\boldsymbol{X}_i}, \overline{\boldsymbol{B}} - \overline{\boldsymbol{B}}^* \rangle + (\beta - \beta^*) \right| \boldsymbol{u}_i^{\mathsf{T}} \overline{\boldsymbol{X}_i} \boldsymbol{v}_j \right| \\
&\leq C\sqrt{r'} d_n n^{-M} (I_1I_2 \vee I_2I_3) \log n,
\end{aligned}$$

which implies

$$S_2 \leq C \frac{\sqrt{r'} d_n (I_1I_3 \vee I_2I_3) \log n}{n^M}, \quad \text{(S2.13)}$$

with high probability. Therefore, for sufficiently large M*, together with* (S2.12)

and (S2.13)*,*

$$\sup_{(\beta,\overline{\boldsymbol{B}})\in\mathscr{A}'}\left\|\frac{1}{n}\sum_{i=1}^{n}q_i(\beta,\overline{\boldsymbol{B}})\right\|_{op}\leq Cd_n\sqrt{\frac{r'(I_1I_3\vee I_2I_3)\log n}{n}}+Cd_n\frac{(r')^{3/2}(I_1I_3\vee I_2I_3)^2(\log n)^2}{n},$$

with probability at least $1-n^{-C}$*, which yields*

$$\begin{aligned}&\sup_{(\beta,\boldsymbol{\mathcal{B}})\in\mathscr{A}}\left\|\frac{1}{n}\sum_{i=1}^{n}(L_h'(u_i)-L_h'(u_i^*))y_i\boldsymbol{\mathcal{X}}_i-\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(L_h'(u_i)-L_h'(u_i^*))y_i\boldsymbol{\mathcal{X}}_i]\right\|\\&\leq Cd_n\sqrt{\frac{r(I_1I_3\vee I_2I_3)\log n}{n}}+Cd_n\frac{r^{3/2}(I_1I_3\vee I_2I_3)^2(\log n)^2}{n}.\end{aligned}$$

■

Proof of Lemma 1: *We first prove* (2.6)*. By the optimality of* $\boldsymbol{\mathcal{B}}^{k+1}$*,*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}L_h(u_i^k)+\frac{1}{n}\sum_{i=1}^{n}L_h'(u_i^k)y_i\langle\boldsymbol{\mathcal{B}}^{k+1}-\boldsymbol{\mathcal{B}}^k,\boldsymbol{\mathcal{X}}_i\rangle+\frac{\mu}{2}\|\boldsymbol{\mathcal{B}}^{k+1}-\boldsymbol{\mathcal{B}}^k\|_F^2+\lambda\|\boldsymbol{\mathcal{B}}^{k+1}\|_*\\&\leq\frac{1}{n}\sum_{i=1}^{n}L_h(u_i^k)+\lambda\|\boldsymbol{\mathcal{B}}^k\|_*.\end{aligned}\tag{S2.14}$$

On the other hand, by the descent lemma (Beck and Teboulle, 2009, Lemma 2.1),

we get

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}L_h(y_i(\beta^k+\langle\boldsymbol{\mathcal{B}}^{k+1},\boldsymbol{\mathcal{X}}_i\rangle))\\&\leq\frac{1}{n}\sum_{i=1}^{n}L_h(u_i^k)+\frac{1}{n}\sum_{i=1}^{n}L_h'(u_i^k)y_i\langle\boldsymbol{\mathcal{B}}^{k+1}-\boldsymbol{\mathcal{B}}^k,\boldsymbol{\mathcal{X}}_i\rangle+\frac{C_{L_h}}{2}\|\boldsymbol{\mathcal{B}}^{k+1}-\boldsymbol{\mathcal{B}}^k\|_F^2.\end{aligned}\tag{S2.15}$$

Adding (S2.14) *and* (S2.15)*, it can be seen that*

$$\frac{1}{n}\sum_{i=1}^{n} L_h(y_i(\beta^k + \langle \boldsymbol{\mathcal{B}}^{k+1}, \boldsymbol{\mathcal{X}}_i\rangle)) + \lambda\|\boldsymbol{\mathcal{B}}^{k+1}\|_* + \frac{\mu - C_{L_h}}{2}\|\boldsymbol{\mathcal{B}}^{k+1} - \boldsymbol{\mathcal{B}}^k\|_F^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n} L_h(u_i^k) + \lambda\|\boldsymbol{\mathcal{B}}^k\|_*,$$

which leads to (2.6)*. Then we prove* (2.7)*. Note that*

$$f(\beta^k, \boldsymbol{\mathcal{B}}^{k+1}) - f(\beta^{k+1}, \boldsymbol{\mathcal{B}}^{k+1})$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_h(y_i(\beta^k + \langle \boldsymbol{\mathcal{B}}^{k+1}, \boldsymbol{\mathcal{X}}_i\rangle)) - \frac{1}{n}\sum_{i=1}^{n} L_h(y_i(\beta^{k+1} + \langle \boldsymbol{\mathcal{B}}^{k+1}, \boldsymbol{\mathcal{X}}_i\rangle)).$$

By the iteration equation of β^{k+1}*,*

$$-\frac{1}{n}\sum_{i=1}^{n} L_h'(y_i(\beta^k + \langle \boldsymbol{\mathcal{B}}^{k+1}, \boldsymbol{\mathcal{X}}_i\rangle))y_i = \frac{1}{\rho}(\beta^{k+1} - \beta^k),$$

which implies, with the descent lemma,

$$f(\beta^k, \boldsymbol{\mathcal{B}}^{k+1}) - f(\beta^{k+1}, \boldsymbol{\mathcal{B}}^{k+1}) \geq (\frac{1}{\rho} - \frac{C_{L_h}}{2})|\beta^{k+1} - \beta^k|^2 = \gamma_2|\beta^{k+1} - \beta^k|^2.$$

■

Proof of Theorem 3: (1). *By Lemma 1 and the condition on γ, we have*

$$f(\boldsymbol{\mathcal{W}}^k) - f(\boldsymbol{\mathcal{W}}^{k+1}) \geq \gamma\{\|\boldsymbol{\mathcal{B}}^{k+1} - \boldsymbol{\mathcal{B}}^k\|_F^2 + |\beta^{k+1} - \beta^k|^2\} = \gamma\|\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{W}}^k\|_F^2, \tag{S2.16}$$

which means $f(\boldsymbol{\mathcal{W}}^k)$ is always monotonically decreasing.

(2). *Note that $f(\boldsymbol{\mathcal{W}}^k)$ is bounded. Then, the sequence $\{\boldsymbol{\mathcal{W}}^k\}$ is also bounded via the coerciveness of $f(\cdot)$ (the coerciveness of a function refers to its property of tending toward infinity as the input variables approach infinity, i.e., a function $g(x) \to +\infty$ as $\|x\| \to +\infty$ where $\|x\|$ denotes the norm (e.g., Euclidean norm) of x). Summing up* (S2.16) *for all $k > 0$, we immediately have*

$$f(\boldsymbol{\mathcal{W}}^0) \geq \gamma \sum_{k=0}^{+\infty} \|\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{W}}^k\|_F,$$

which implies

$$\sum_{k=0}^{+\infty} \|\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{W}}^k\|_F \leq \frac{1}{\gamma} f(\boldsymbol{\mathcal{W}}^0) < +\infty,$$

Hence, we get

$$\lim_{k\to+\infty}(\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{W}}^k) = 0.$$

(3). *We only prove the accumulation point $\boldsymbol{\mathcal{B}}^\star$ is a critical point of $f(\beta, \boldsymbol{\mathcal{B}})$ since the same assertion of β is easy to obtain. By the boundedness of $\{\boldsymbol{\mathcal{B}}^k\}$, we*

have a subsequence $\{\boldsymbol{\mathcal{B}}^{k_t}\}$ *satisfying* $\lim_{t\to+\infty}\boldsymbol{\mathcal{B}}^{k_t} = \boldsymbol{\mathcal{B}}^\star$*, and* $\{\boldsymbol{\mathcal{B}}^{k_t}\}$ *is bounded. Due to the optimality condition of* $\boldsymbol{\mathcal{B}}^{k_t+1}$*, there exists* $\boldsymbol{\mathcal{Z}}^{k_t+1} \in \partial\|\boldsymbol{\mathcal{B}}^{k_t+1}\|_*$ *such that*

$$\lambda\boldsymbol{\mathcal{Z}}^{k_t+1} + \frac{1}{n}\sum_{i=1}^{n} L_h'(u_i^{k_t})y_i\boldsymbol{\mathcal{X}}_i + \mu(\boldsymbol{\mathcal{B}}^{k_t+1} - \boldsymbol{\mathcal{B}}^{k_t}) = 0, \tag{S2.17}$$

where $\partial\|\boldsymbol{\mathcal{B}}\|_*$ *indicates the subdifferential of t-TNN. Let* $t \to +\infty$ *in* (S2.17)*, then there exists* $\boldsymbol{\mathcal{Z}}^\star \in \partial\|\boldsymbol{\mathcal{B}}^\star\|_*$ *such that*

$$0 = \lambda\boldsymbol{\mathcal{Z}}^\star + \frac{1}{n}\sum_{i=1}^{n} L_h'(u_i^\star)y_i\boldsymbol{\mathcal{X}}_i \in \partial_{\boldsymbol{\mathcal{B}}} f(\boldsymbol{\mathcal{W}}^\star),$$

which completes the proof. ■

S3 Proofs and Lemmas for results in Section 3

Lemma S3.1 *Suppose* ***Assumption*** *1-5 hold, with probability* $1 - n^{-C}$*, we have*

$$\begin{aligned}\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}} \big|&Q_1(\beta,\boldsymbol{\mathcal{B}}) - Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*) - \delta\nabla_\beta Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*) - \langle\boldsymbol{\mathcal{D}}, \nabla_{\boldsymbol{\mathcal{B}}} Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*)\rangle \\ &- \mathbb{E}[L_h(y_i(\beta + \langle\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{X}}_i\rangle))] + \mathbb{E}[L_h(y_i(\beta^* + \langle\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{X}}_i\rangle))]\big| \\ &\le C\Big(\frac{b_n\sqrt{r}tI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}t)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big).\end{aligned}$$

Proof of Lemma S3.1: *With the definition of* $Q_1(\cdot,\cdot)$*, for brevity, let* $f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}}) = L_h(u_i) - L_h(u_i^*) - \delta L_h'(u_i^*)y_i - L_h'(u_i^*)y_i\langle\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{X}}_i\rangle$*. We restrict to the event*

$\max_{1\le i\le n}\|\boldsymbol{\mathcal{X}}_i\| \le b_n := C\sqrt{(I_1I_2 \vee I_2I_3)\log n}$. *By a direct computation, we have*

$$|f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}})| \le C|\delta + \langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle| \le Cb_n\sqrt{r}t,$$

and

$$f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}}) = L_h''(u_i^*)y_i^2\delta^2 + 2L_h''(u_i^*)y_i^2\delta\langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle + L_h''(u_i^*)y_i^2\langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle^2 + o((\delta+\langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle)^2).$$

Thus,

$$\mathbb{E}[f^2(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}})] \le Cb_n\sqrt{r}t\cdot\{\mathbb{E}[(\delta+\langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle)^2] + o((\delta+\langle \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{X}}_i\rangle)^2)\} \le C(b_n\sqrt{r}t)^3.$$

For any fixed $(\delta,\boldsymbol{\mathcal{D}})$*, applying Bernstein's inequality yields*

$$P(|(P_n - P)f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}})| > u) \le exp\Big\{-\frac{Cnu^2}{ub_n\sqrt{r}t + (b_n\sqrt{r}t)^3}\Big\}. \quad \text{(S3.18)}$$

To make the bound uniformly over $\mathscr{S} := \{(\delta,\boldsymbol{\mathcal{D}}) : |\delta| + \|\boldsymbol{\mathcal{D}}\|_* \le C\sqrt{r}t\}$*, we resort to the covering argument. Suppose* $\{(\delta_1,\boldsymbol{\mathcal{D}}_1),\cdots,(\delta_M,\boldsymbol{\mathcal{D}}_M)\}$ *is* ζ*-covering of* $\mathscr{S}$ *with* $M \le (C/\zeta)^{I_1I_2I_3^2}$*. It is obvious that* $f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}})$ *is Lipschitz continuous. Then, by choosing* $\zeta \asymp n^{-C}$ *and taking* $u = C\big(\frac{b_n\sqrt{r}tI_1I_2I_3^2\log n}{n} +$

$\frac{(b_n\sqrt{r}t)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\big)$ *in* (S3.18)*, we have*

$$\sup_{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{T}}|(P_n-P)f(\boldsymbol{\mathcal{X}}_i,\delta,\boldsymbol{\mathcal{D}})|\le C\Big(\frac{b_n\sqrt{r}tI_1I_2I_3^2\log n}{n}+\frac{(b_n\sqrt{r}t)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big).$$

with probability at least $1-n^{-C}$. ■

Lemma S3.2 *Suppose* ***Assumption*** *1-5 hold, with probability* $1-n^{-C}$*, we have*

$$\begin{aligned}&|\nabla_\beta\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)|+\|\nabla_{\boldsymbol{\mathcal{B}}}\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)\|\\&\le C\sqrt{\frac{(I_1I_3\vee I_2I_3)\log N}{N}}+C\Big(\frac{r(I_1I_3\vee I_2I_3)}{n}\sqrt{\log n}+\frac{r^2(I_1I_3\vee I_2I_3)^{5/2}(\log n)^2}{n^{3/2}}\Big).\end{aligned}$$

Proof of Lemma S3.2: *We only consider* $\|\nabla_{\boldsymbol{\mathcal{B}}}\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)\|$ *because the bound of* $|\nabla_\beta\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)|$ *can be derived in the same way. By the definition of* $\breve{Q}(\cdot,\cdot)$*, it leads to* $\nabla_{\boldsymbol{\mathcal{B}}}\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)=\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*)-\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})+\nabla_{\boldsymbol{\mathcal{B}}}Q(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})$*, and thus*

$$\begin{aligned}\|\nabla_{\boldsymbol{\mathcal{B}}}\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)\|\le&\|\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*)-\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})-\nabla_{\boldsymbol{\mathcal{B}}}Q(\beta^*,\boldsymbol{\mathcal{B}}^*)\\&+\nabla_{\boldsymbol{\mathcal{B}}}Q(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})\|+\|\nabla_{\boldsymbol{\mathcal{B}}}Q(\beta^*,\boldsymbol{\mathcal{B}}^*)\|.\end{aligned}$$

By a similar argument as Lemma S2.4,

$$\|\nabla_{\boldsymbol{\mathcal{B}}}Q(\beta^*,\boldsymbol{\mathcal{B}}^*)\|\le C\sqrt{\frac{(I_1I_3\vee I_2I_3)\log N}{N}}.$$

Further, applying Lemma S2.5 twice yields

$$\begin{aligned}&\|\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\beta^*,\boldsymbol{\mathcal{B}}^*)-\nabla_{\boldsymbol{\mathcal{B}}}Q_1(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})-\nabla_{\boldsymbol{\mathcal{B}}}Q(\beta^*,\boldsymbol{\mathcal{B}}^*)+\nabla_{\boldsymbol{\mathcal{B}}}Q(\widehat{\beta},\widehat{\boldsymbol{\mathcal{B}}})\|\\&\leq C\Big(\frac{r(I_1I_3\vee I_2I_3)}{n}\sqrt{\log n}+\frac{r^2(I_1I_3\vee I_2I_3)^{5/2}(\log n)^2}{n^{3/2}}\Big).\end{aligned}$$

Integrating the above two bounds completes the proof. ■

Proof of Theorem 4: *By direct algebraic computation,*

$$\breve{Q}(\beta,\boldsymbol{\mathcal{B}})-\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)\geq\nabla\breve{Q}_\beta(\beta^*,\boldsymbol{\mathcal{B}}^*)(\beta-\beta^*)+\langle\breve{Q}_{\boldsymbol{\mathcal{B}}}(\beta^*,\boldsymbol{\mathcal{B}}^*),\boldsymbol{\mathcal{B}}-\boldsymbol{\mathcal{B}}^*\rangle.$$

The following proof is similar to that in Theorem 1. We have $(\breve{\delta},\breve{\boldsymbol{\mathcal{D}}})=(\breve{\beta}-\beta^*,\breve{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}^*)\in\mathscr{H}:=\{(\delta,\boldsymbol{\mathcal{D}}):\|\boldsymbol{\mathcal{D}}^{\perp}\|_*\leq 3\|\boldsymbol{\mathcal{D}}^0\|_*+|\delta|\}$ *if* $\lambda\geq 2\|\breve{Q}_{\boldsymbol{\mathcal{B}}}(\beta^*,\boldsymbol{\mathcal{B}}^*)\|$ *and* $\lambda\geq 2|\nabla\breve{Q}_\beta(\beta^*,\boldsymbol{\mathcal{B}}^*)|$*, which has been completed in Lemma S3.2 with* λ *defined in the statement of the theorem. Next, we will connect* $\breve{Q}(\beta,\boldsymbol{\mathcal{B}})-\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)$ *to* $\mathbb{E}[(\beta-\beta^*+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{B}}-\boldsymbol{\mathcal{B}}^*\rangle)^2]$ *to bound the difference between empirical risk and population risk corresponding to the central estimator. Assume* $\mathbb{E}[(\breve{\delta},\langle\boldsymbol{\mathcal{X}},\breve{\boldsymbol{\mathcal{D}}}\rangle)^2]\geq(La_n)^2$ *with* a_n *in theorem and a sufficiently large* $L>0$*. As in the proof of Theorem 1, this implies*

$$\inf_{\substack{(\delta,\boldsymbol{\mathcal{D}})\in\mathscr{H}\\ \mathbb{E}[(\delta+\langle\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{D}}\rangle)^2]\geq(La_n)^2}}\breve{Q}(\beta^*+\delta,\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}})-\breve{Q}(\beta^*,\boldsymbol{\mathcal{B}}^*)+\lambda\|\boldsymbol{\mathcal{B}}^*+\boldsymbol{\mathcal{D}}\|_*-\lambda\|\boldsymbol{\mathcal{B}}^*\|<0.$$

Then, for some $(\delta,\boldsymbol{\mathcal{D}})$ *that makes the above expression negative, using* $|\delta|+$

$\|\boldsymbol{\mathcal{D}}\|_* \le C\sqrt{r}\|\boldsymbol{\mathcal{D}}\|_F + C|\delta| \le C\sqrt{r}La_n$ and Lemma S3.1 yields

$$
\begin{aligned}
C(La_n)^2 &\le \mathbb{E}[L_h(y_i(\beta + \langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{X}}_i\rangle))] - \mathbb{E}[L_h(y_i(\beta^* + \langle \boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}}_i\rangle))] \\
&\le \breve{Q}(\beta, \boldsymbol{\mathcal{B}}) - \breve{Q}(\beta^*, \boldsymbol{\mathcal{B}}^*) - \nabla\breve{Q}_\beta(\beta^*, \boldsymbol{\mathcal{B}}^*)(\beta - \beta^*) - \langle \breve{Q}_{\boldsymbol{\mathcal{B}}}(\beta^*, \boldsymbol{\mathcal{B}}^*), \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{B}}^*\rangle \\
&\quad + C\Big(\frac{b_n\sqrt{r}La_nI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}La_n)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big) \\
&\le \breve{Q}(\beta^* + \delta, \boldsymbol{\mathcal{B}}^* + \boldsymbol{\mathcal{D}}) - \breve{Q}(\beta^*, \boldsymbol{\mathcal{B}}^*) + C\lambda(|\delta| + \|\boldsymbol{\mathcal{D}}\|_*) \\
&\quad + C\Big(\frac{b_n\sqrt{r}La_nI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}La_n)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big) \\
&\le \lambda\|\boldsymbol{\mathcal{B}}^*\|_* - \lambda\|\boldsymbol{\mathcal{B}}^* + \boldsymbol{\mathcal{D}}\|_* + C\lambda(|\delta| + \|\boldsymbol{\mathcal{D}}\|_*) \\
&\quad + C\Big(\frac{b_n\sqrt{r}La_nI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}La_n)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big) \\
&\le \lambda\|\boldsymbol{\mathcal{D}}\|_* + C\lambda(|\delta| + \|\boldsymbol{\mathcal{D}}\|_*) \\
&\quad + C\Big(\frac{b_n\sqrt{r}La_nI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}La_n)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big) \\
&\le C\lambda\sqrt{r}La_n + C\Big(\frac{b_n\sqrt{r}La_nI_1I_2I_3^2\log n}{n} + \frac{(b_n\sqrt{r}La_n)^{3/2}\sqrt{I_1I_2I_3^2\log n}}{\sqrt{n}}\Big) \\
&\le CLa_n^2 + CL^{3/2}a_n^2,
\end{aligned}
$$

where the third inequality uses Hölder's inequality and Lemma S3.2. The above displayed leads to a contradiction when L is large enough. Thus we have $\mathbb{E}[(\breve{\delta} + \langle \boldsymbol{\mathcal{X}}, \breve{\boldsymbol{\mathcal{D}}}\rangle)^2] \le (La_n^2)^2$ and the rest of the statements can be obtained easily as for Theorem 1.

Proof of Theorem 5: *There exists $\boldsymbol{\mathcal{Z}} \in \partial\|\breve{\boldsymbol{\mathcal{B}}}\|_*$ by the first order optimality con-*

dition. Then, we have

$$\frac{1}{n}\sum_{i=1}^{n} L_h'(\breve{v}_i)y_i\overline{\boldsymbol{X}_i} - \frac{1}{n}\sum_{i=1}^{n} L_h'(\hat{v}_i)y_i\overline{\boldsymbol{X}_i} + \frac{1}{n}\sum_{i=1}^{N} L_h'(\hat{v}_i)y_i\overline{\boldsymbol{X}_i} + \lambda\overline{\boldsymbol{Z}} = 0.$$

Let $\mathrm{rank_t}(\breve{\boldsymbol{\mathcal{B}}}) = \breve{r}$*, we have* $\mathrm{rank}(\overline{\breve{\boldsymbol{B}}}) = \breve{r}' \leq I_3\breve{r}$ *and its SVD given by* $\overline{\breve{\boldsymbol{B}}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$*,* $\boldsymbol{U} \in \mathbb{R}^{I_1I_3\times\hat{r}'}, \boldsymbol{V} \in \mathbb{R}^{I_2I_3\times\breve{r}'}$*. Using (2.1) in Koltchinskii et al. (2011), we have*

$$\partial\|\overline{\breve{\boldsymbol{B}}}\|_* = \{\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}} + \boldsymbol{U}^{\perp}\boldsymbol{W}(\boldsymbol{V}^{\perp})^{\mathsf{T}} : \|\boldsymbol{W}\|_{op} \leq 1\},$$

where $\boldsymbol{U}^{\perp} \in \mathbb{R}^{I_1I_3\times(I_1I_3-\hat{r}')}$ *and* $\boldsymbol{V}^{\perp} \in \mathbb{R}^{I_2I_3\times(I_2I_3-\breve{r}')}$ *are orthogonal matrices and their columns are orthogonal to those in* $\boldsymbol{U}$ *and* $\boldsymbol{V}$*, respectively. Let* $\boldsymbol{u}_j, \boldsymbol{v}_j, j = 1, \cdots, \hat{r}'$ *be the columns of* $\boldsymbol{U}$ *and* $\boldsymbol{V}$*. Pre(post)-multiplying* $\boldsymbol{u}_j$ *and* $\boldsymbol{v}_j$ *to* (S2.10) *yields,*

$$\begin{aligned}-\lambda &= \frac{1}{n}\sum_{i=1}^{n} L_h'(\breve{v}_i)y_i\boldsymbol{u}_j^{\mathsf{T}}\overline{\boldsymbol{X}_i}\boldsymbol{v}_j - \frac{1}{n}\sum_{i=1}^{n} L_h'(\hat{v}_i)y_i\boldsymbol{u}_j^{\mathsf{T}}\overline{\boldsymbol{X}_i}\boldsymbol{v}_j + \frac{1}{n}\sum_{i=1}^{N} L_h'(\hat{v}_i)y_i\boldsymbol{u}_j^{\mathsf{T}}\overline{\boldsymbol{X}_i}\boldsymbol{v}_j \\ &:= g_j,\ j = 1, \cdots, \breve{r}'.\end{aligned}$$

By the similar argument as that in Theorem 2 and Lemma S2.5, we have $\breve{r} \leq Cr$*.* ■

Table S4.1: Prediction errors (%) of different estimators with $N = 10000$ and different values of m.

r	d	m	CSL		Ave		Sub	
			Gaussian	Epanechnikov	Gaussian	Epanechnikov	Gaussian	Epanechnikov
2	0.5	1	14.97 (0.0034)	13.44 (0.0031)	14.97 (0.0034)	13.44 (0.0031)	14.97 (0.0034)	13.44 (0.0031)
		5	14.46 (0.0172)	12.60 (0.0084)	16.35 (0.0094)	15.98 (0.0077)	18.44 (0.0099)	17.64 (0.0084)
		10	14.79 (0.0195)	13.89 (0.0161)	17.96 (0.0218)	17.43 (0.0148)	21.73 (0.0257)	21.06 (0.0193)
		20	15.14 (0.0406)	14.08 (0.0369)	21.61 (0.0194)	20.01 (0.0161)	27.67 (0.0552)	26.68 (0.0579)
	-1	1	7.53 (0.0032)	6.05 (0.0027)	7.53 (0.0032)	6.05 (0.0027)	7.53 (0.0032)	6.05 (0.0027)
		5	5.04 (0.0074)	4.78 (0.0077)	7.72 (0.0071)	6.49 (0.0079)	7.79 (0.0071)	7.78 (0.0082)
		10	5.50 (0.0154)	5.39 (0.0187)	9.04 (0.0175)	8.16 (0.0183)	10.36 (0.0223)	10.14 (0.0253)
		20	6.32 (0.0224)	6.30 (0.0152)	11.02 (0.0173)	10.96 (0.0194)	15.24 (0.0379)	14.84 (0.0330)
5	-0.5	1	16.83 (0.0083)	14.97 (0.0024)	16.83 (0.0083)	14.97 (0.0024)	16.83 (0.0083)	14.97 (0.0024)
		5	15.12 (0.0120)	13.50 (0.0144)	21.98 (0.0184)	19.19 (0.0179)	23.36 (0.0185)	21.89 (0.0100)
		10	16.67 (0.0203)	14.95 (0.0136)	25.57 (0.0307)	24.45 (0.0237)	28.92 (0.0336)	27.80 (0.0227)
		20	17.48 (0.0453)	15.32 (0.0427)	31.31 (0.0394)	29.63 (0.0141)	37.28 (0.0898)	37.82 (0.0797)
	-1	1	6.94 (0.0027)	6.29 (0.0062)	6.94 (0.0027)	6.29 (0.0062)	6.94 (0.0027)	6.29 (0.0062)
		5	5.94 (0.0061)	5.15 (0.0104)	10.31 (0.0104)	10.24 (0.0109)	11.48 (0.0073)	11.75 (0.0073)
		10	6.54 (0.0109)	6.31 (0.0085)	14.55 (0.0099)	13.92 (0.0197)	15.31 (0.0183)	16.16 (0.0180)
		20	7.70 (0.0114)	7.26 (0.0107)	18.65 (0.0296)	18.48 (0.0145)	21.28 (0.0377)	23.12 (0.0583)

Table S4.2: Prediction errors of different estimators with $n = 1500$ and different values of m.

r	d	m	CSL		Ave		Sub	
			Gaussian	Epanechnikov	Gaussian	Epanechnikov	Gaussian	Epanechnikov
2	-0.5	1	18.46 (0.0137)	20.42 (0.0200)	18.46 (0.0137)	20.42 (0.0200)	18.46 (0.0137)	20.42 (0.0200)
		5	11.20 (0.0131)	12.61 (0.0153)	13.26 (0.0173)	13.37 (0.0153)	18.29 (0.0154)	20.60 (0.0157)
		10	9.91 (0.0169)	11.69 (0.0250)	12.02 (0.0134)	12.94 (0.0185)	18.55 (0.0101)	20.38 (0.0206)
		20	10.24 (0.0224)	11.6 (0.0192)	11.44 (0.0198)	11.90 (0.0210)	18.40 (0.0215)	20.11 (0.0241)
	-1	1	7.64 (0.0047)	9.54 (0.0093)	7.64 (0.0047)	9.54 (0.0093)	7.64 (0.0047)	9.54 (0.0093)
		5	3.54 (0.0066)	4.27 (0.0071)	6.34 (0.0073)	6.79 (0.0081)	7.92 (0.0053)	9.40 (0.0114)
		10	3.38 (0.0101)	3.43 (0.0098)	6.49 (0.0109)	7.02 (0.0122)	7.78 (0.0082)	9.72 (0.0129)
		20	2.70 (0.0057)	3.8 (0.0085)	5.96 (0.0061)	6.03 (0.0088)	8.12 (0.0069)	9.22 (0.0115)
5	-0.5	1	25.42 (0.0203)	27.28 (0.0200)	25.42 (0.0203)	27.28 (0.0200)	25.42 (0.0203)	27.28 (0.0200)
		5	11.23 (0.0205)	11.37 (0.0354)	16.63 (0.0211)	17.60 (0.0382)	25.26 (0.0200)	27.52 (0.0392)
		10	10.44 (0.0365)	10.85 (0.0242)	15.49 (0.0225)	16.61 (0.0265)	25.10 (0.0210)	27.47 (0.0258)
		20	9.88 (0.0266)	10.21 (0.0227)	14.95 (0.0237)	15.31 (0.0249)	25.16 (0.0251)	27.24 (0.0218)
	-1	1	12.15 (0.0125)	14.82 (0.0182)	12.15 (0.0125)	14.82 (0.0182)	12.15 (0.0125)	14.82 (0.0182)
		5	3.95 (0.0121)	3.56 (0.0094)	9.57 (0.0137)	9.92 (0.0105)	11.86 (0.0091)	14.39 (0.0162)
		10	3.23 (0.0075)	2.69 (0.0075)	9.07 (0.0089)	9.46 (0.0085)	11.89 (0.0067)	14.73 (0.0165)
		20	2.72 (0.0072)	2.66 (0.0045)	8.57 (0.0078)	8.71 (0.0096)	12.17 (0.0118)	14.55 (0.0149)

S4 Some results of simulations

S5 Extensions to order-d tensors

As we mentioned in Section 2, our results can be extended to the order-d tensor. In this section, we briefly state the key ideas by introducing the order-d t-product.

For a given high-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$, the block circulant operator

Table S4.3: Prediction errors of different estimators with $n = 1500$, $m = 10$ and different values of I.

r	d	I	CSL		Ave		Sub	
			Gaussian	Epanechnikov	Gaussian	Epanechnikov	Gaussian	Epanechnikov
2	-0.5	20	9.91 (0.0169)	11.69 (0.0250)	17.32 (0.0106)	19.03 (0.0219)	18.55 (0.0101)	20.38 (0.0206)
		40	11.62 (0.0321)	13.84 (0.0543)	19.28 (0.0152)	21.32 (0.0207)	20.018 (0.0139)	22.89 (0.0237)
		60	12.73 (0.0510)	14.23 (0.0502)	21.61 (0.0193)	23.69 (0.0348)	22.56 (0.0243)	24.20 (0.0317)
		100	16.79 (0.0855)	18.24 (0.0764)	25.90 (0.0257)	27.51 (0.0294)	26.43 (0.0281)	28.41 (0.0318)
		200	20.33 (0.0204)	19.52 (0.0114)	32.21 (0.0943)	33.67 (0.0318)	35.45 (0.1253)	34.48 (0.0254)
	1	20	3.38 (0.0101)	3.43 (0.0098)	7.15 (0.0104)	7.73 (0.0146)	7.78 (0.0082)	9.72 (0.0129)
		40	4.12 (0.0167)	4.57 (0.0100)	9.01 (0.0114)	10.02 (0.0135)	9.38 (0.0110)	11.74 (0.0123)
		60	5.23 (0.0105)	6.23 (0.0143)	11.73 (0.0106)	13.90 (0.0185)	11.94 (0.0089)	14.19 (0.018)
		100	8.43 (0.0321)	8.77 (0.0255)	15.84 (0.0384)	18.10 (0.0419)	16.64 (0.0565)	19.84 (0.0448)
		200	11.92 (0.0434)	11.71 (0.0108)	24.69 (0.0481)	26.81 (0.4326)	26.03 (0.0257)	27.57 (0.0304)
5	-0.5	20	10.44 (0.0365)	10.85 (0.0242)	24.80 (0.0143)	24.95 (0.0164)	25.10 (0.0210)	27.47 (0.0258)
		40	11.72 (0.0247)	11.91 (0.0409)	26.91 (0.0344)	26.71 (0.0272)	27.37 (0.0261)	29.71 (0.0312)
		60	13.01 (0.0142)	13.08 (0.0223)	28.51 (0.0667)	28.94 (0.0331)	29.51 (0.0235)	31.91 (0.0428)
		100	16.63 (0.1101)	16.78 (0.1306)	31.76 (0.0285)	33.08 (0.0271)	33.18 (0.0609)	34.22 (0.0639)
		200	19.09 (0.3309)	18.99 (0.4922)	39.81 (0.0374)	40.64 (0.2074)	41.90 (0.1445)	42.86 (0.1918)
	1	20	3.23 (0.0075)	2.69 (0.0075)	11.25 (0.0077)	14.05 (0.0164)	11.89 (0.0067)	14.73 (0.0165)
		40	4.36 (0.0054)	3.62 (0.0049)	13.22 (0.0162)	15.37 (0.0382)	13.87 (0.0208)	16.28 (0.0186)
		60	5.38 (0.0048)	4.41 (0.0083)	14.53 (0.0203)	16.7 (0.0135)	15.67 (0.0162)	17.48 (0.0504)
		100	9.5 (0.018)	8.94 (0.0168)	19.13 (0.0188)	20.54 (0.0275)	20.14 (0.0285)	22.64 (0.0366)
		200	15.70 (0.022)	15.10 (0.0149)	27.65 (0.0702)	29.48 (0.0615)	30.61 (0.0690)	32.86 (0.0637)

Table S4.4: Rank estimation of different estimators with $n = 1500$, $m = 10$ and different values of I.

r	d	I	CSL		Ave		Sub	
			Gaussian	Epanechnikov	Gaussian	Epanechnikov	Gaussian	Epanechnikov
2	-0.5	20	2.30 (0.5408)	2.02 (0.6224)	2.40 (0.6135)	2.20 (0.3448)	2.36 (0.4800)	2.10 (0.3265)
		40	1.96 (0.6351)	1.94 (0.7514)	2.22 (0.4463)	1.76 (0.6853)	2.16 (0.4249)	1.88 (0.2894)
		60	2.02 (0.5996)	2 (0.6531)	1.98 (0.4992)	2.12 (0.2971)	1.96 (0.3873)	2.20 (0.4490)
		100	1.94 (0.8739)	1.98 (0.6792)	2.14 (0.3732)	1.88 (0.3367)	1.88 (0.8462)	2.14 (0.2698)
		200	1.54 (0.6176)	1.34 (0.8820)	1 (0)	1 (0)	1 (0)	1 (0)
	1	20	2.02 (0.7347)	2.02 (0.3873)	2.32 (0.4751)	2.08 (0.3184)	2.08 (0.4016)	2.1 (0.2959)
		40	2.08 (0.6955)	2.06 (0.518)	1.92 (0.4637)	2.12 (0.3491)	2.12 (0.4392)	1.94 (0.5065)
		60	1.96 (0.5310)	2.18 (0.5936)	1.88 (0.5934)	2.22 (0.4715)	1.92 (0.4490)	2.34 (0.3139)
		100	2.08 (0.8506)	1.96 (0.6147)	2.08 (0.3349)	2.14 (0.6947)	2.16 (0.6327)	2.16 (0.3412)
		200	1.63 (0.6174)	1.5 (0.6231)	1 (0)	1 (0)	1 (0)	1 (0)
5	-0.5	20	5 (0.6122)	5.04 (0.4065)	4.78 (0.4632)	4.88 (0.4641)	4.56 (0.3331)	5.14 (0.4494)
		40	4.96 (0.4882)	5.04 (0.6045)	5.12 (0.4673)	4.92 (0.7639)	5.54 (0.5392)	4.84 (0.6106)
		60	4.88 (0.4082)	5.08 (0.4914)	4.98 (0.7614)	5.08 (0.6431)	5.18 (0.2731)	4.92 (0.3269)
		100	4.98 (0.7106)	4.96 (1.0600)	5.06 (0.7616)	5.14 (0.5273)	4.94 (0.6149)	4.86 (0.8555)
		200	4.18 (0.9130)	4.74 (0.9420)	1.96 (0.6427)	1.88 (0.8466)	1.94 (0.1567)	1.92 (0.1982)
	1	20	5.18 (0.3955)	5.02 (0.4282)	5.02 (0.4672)	5.08 (0.6416)	4.56 (0.2922)	4.98 (0.4282)
		40	5 (0.4868)	4.98 (0.6224)	4.88 (0.5271)	5.12 (0.9461)	5.26 (0.3596)	5.06 (0.4392)
		60	5.02 (0.6302)	5.02 (0.5506)	4.96 (0.2597)	4.98 (0.4157)	5.32 (0.6302)	5.18 (0.4812)
		100	5.12 (0.5669)	5.34 (0.3922)	5.12 (0.4634)	4.82 (0.4483)	4.88 (0.0200)	4.32 (0.1016)
		200	2.57 (0.2857)	2.74 (0.4412)	1.88 (0.8196)	1.76 (0.6582)	2 (0)	1.68 (0.7310)

$\mathrm{circ}(\cdot)$ is defined on the d-th index of tensor $\boldsymbol{\mathcal{A}}$ as

$$\mathrm{circ}(\boldsymbol{\mathcal{A}}) := \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 & \boldsymbol{\mathcal{A}}_d & \cdots & \boldsymbol{\mathcal{A}}_2 \\ \boldsymbol{\mathcal{A}}_2 & \boldsymbol{\mathcal{A}}_1 & \cdots & \boldsymbol{\mathcal{A}}_3 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\mathcal{A}}_d & \boldsymbol{\mathcal{A}}_{d-1} & \cdots & \boldsymbol{\mathcal{A}}_1 \end{bmatrix},$$

which is of size $I_1 I_d \times I_2 I_d \times \cdots \times I_{d-2} I_d \times I_{d-1}$. A block unfolding operator $\mathrm{bvec}(\cdot)$, along with its corresponding inverse operator $\mathrm{bvfold}(\cdot)$, functions on the d-th index of a high-order tensor, are defined as

$$\mathrm{bvec}(\boldsymbol{\mathcal{A}}) := [\boldsymbol{\mathcal{A}}_1, \cdots, \boldsymbol{\mathcal{A}}_d]^{\mathsf{T}}, \ \ \mathrm{bvfold}(\mathrm{bvec}(\boldsymbol{\mathcal{A}})) := \boldsymbol{\mathcal{A}},$$

where $\mathrm{bvec}(\cdot)$ maps $\boldsymbol{\mathcal{A}}$ to an order-$(d-1)$ tensor of size $I_1 I_d \times I_2 \times \cdots \times I_{d-1}$. Thus, the order-$d$ t-product (Martin et al., 2013) is a recursive generalization from the order-3 t-product (Kilmer and Martin, 2011).

Definition S5.1 *Martin et al. (2013)(order-d t-product): The t-product between $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{I_2 \times l \times \cdots \times I_d}$ is defined in a recursive pattern,*

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} := \mathrm{bvfold}(\mathrm{circ}(\boldsymbol{\mathcal{A}}) * \mathrm{bvec}(\boldsymbol{\mathcal{B}})),$$

which is an $I_1 \times l \times \cdots \times I_d$ real tensor.

Specifically, at the base level of recursion, we replace the operator $\mathrm{cric}(\boldsymbol{\mathcal{A}})$ with $\overline{\boldsymbol{A}} = \mathrm{bcirc}(\boldsymbol{\mathcal{A}})$ which is a block circulant matrix of size $I_1 I_3 I_4 \cdots I_d \times I_2 I_3 I_4 \cdots I_d$. Similarly, instead of $\mathrm{bvec}(\boldsymbol{\mathcal{B}})$, we adopt $\mathrm{bunfold}(\boldsymbol{\mathcal{B}})$, which is a block matrix of $I_2 I_3 I_4 \cdots I_d \times l$ generated in the form of recursion from $\boldsymbol{\mathcal{B}}$.

Similarly, by DFT, we have

$$(\boldsymbol{P} \otimes \boldsymbol{I}_{I_1}) \cdot \overline{\boldsymbol{A}} \cdot (\boldsymbol{P}^{\mathsf{H}} \otimes \boldsymbol{I}_{I_2}) = \widetilde{\boldsymbol{A}} = \begin{bmatrix} \widetilde{\boldsymbol{\mathcal{A}}}^{(1)} & & & \\ & \widetilde{\boldsymbol{\mathcal{A}}}^{(2)} & & \\ & & \ddots & \\ & & & \widetilde{\boldsymbol{\mathcal{A}}}^{(J)} \end{bmatrix},$$

where $\boldsymbol{P} = \boldsymbol{F}_{I_d} \otimes \boldsymbol{F}_{I_{d-1}} \otimes \cdots \boldsymbol{F}_{I_3}$, $\boldsymbol{P}^{\mathsf{H}} = \boldsymbol{F}_{I_d}^{\mathsf{H}} \otimes \boldsymbol{F}_{I_{d-1}}^{\mathsf{H}} \otimes \cdots \boldsymbol{F}_{I_3}^{\mathsf{H}}$, $J = I_3 \cdots I_d$ and $\widetilde{\boldsymbol{\mathcal{A}}} = \mathrm{fft}(\boldsymbol{\mathcal{A}}, [], i)$ for $i = 3, \cdots, d$ (using the MATLAB command for ease of exposition). By the unitary invariance, it also yields that,

$$\|\boldsymbol{\mathcal{A}}\|_F = \frac{1}{\sqrt{I_3}} \|\widetilde{\boldsymbol{A}}\|_F, \ \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \frac{1}{I_3} \langle \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{A}} \rangle,$$
$$\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} \Leftrightarrow \widetilde{\boldsymbol{\mathcal{C}}} = \widetilde{\boldsymbol{\mathcal{A}}} \cdot \widetilde{\boldsymbol{\mathcal{B}}}.$$

The order-d conjugate transpose is relatively complex. Since the denifinition

of the orthogonal tensor, the f-diagonal tensor and so on can be easily generalized from Appendix S1.

Definition S5.2 *Qin et al. (2022)(order-d Conjugate transpose): We employ $\boldsymbol{\mathcal{A}}^{\mathsf{H}} \in \mathbb{C}^{I_2 \times I_1 \times \cdots \times I_d}$ to represent the conjugate transpose of $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_d}$ where $\widetilde{\boldsymbol{\mathcal{A}}}^{\mathsf{H}}(:,:,i_3,\cdots,i_d) = \big(\widetilde{\boldsymbol{\mathcal{A}}}(:,:,i_3,\cdots,i_d)\big)^{\mathsf{H}}$ for $i_p \in \{1,\cdots,d_p\}$ with $p \in \{3,\cdots,d\}$.*

With the above arguments of order-d t-product, the t-SVD, tubal rank and t-TNN can also be extended to order-d tensors. We refer the readers to Martin et al. (2013); Qin et al. (2022); Song et al. (2025) for more discussions. Our analysis mainly rely on the operator norm and the decomposability of the t-TNN for which we list the key results which can be proved similarly as Appendix S1.

Definition S5.3 *Define $\|\boldsymbol{\mathcal{A}}\| := \|\overline{\boldsymbol{A}}\|_{op}$ as the tensor spectral norm for $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_d}$.*

Theorem S5.1 *For any order-d tensor $\boldsymbol{\mathcal{A}} \in \mathscr{U}^{\perp}$ and any $\boldsymbol{\mathcal{B}} \in \mathscr{B}$, of the same dimension, with tubal rank r, we have $\|\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}\|_* = \|\boldsymbol{\mathcal{A}}\|_* + \|\boldsymbol{\mathcal{B}}\|_*$, where*

$$\mathscr{U} = \Big\{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{M}} + \boldsymbol{\mathcal{N}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}} : \boldsymbol{\mathcal{M}} \in \mathbb{R}^{r \times I_2 \times \cdots \times I_d}, \boldsymbol{\mathcal{N}} \in \mathbb{R}^{I_1 \times r \times \cdots \times I_d}\Big\},$$
$$\mathscr{B} = \Big\{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^{\mathsf{H}} * \boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{Z}} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}} - \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^{\mathsf{H}} * \boldsymbol{\mathcal{Z}} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^{\mathsf{H}} : \boldsymbol{\mathcal{Z}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}\Big\}.$$

Based on Definition S5.3 and Theorem S5.1, by the same argument as Sections

2 and 3, we can extend the derived convergence rate to the order-d tensor. For example, with the same condition as Corollary 1, we have

$$(\widehat{\beta} - \beta^*)^2 + \|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\|_F^2 \leq O_p\Big(\frac{r(I_1 I_3 I_4 \cdots I_d \vee I_2 I_3 I_4 \cdots I_d)}{n}\Big),$$

for the estimated order-d tensor parameter, where $\widehat{\beta}$ and $\widehat{\boldsymbol{\mathcal{B}}}$ are obtained by the same way as in Section 2. The distributed estimator in Section 3 can be generalized in the same way.

Bibliography

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *Ann. Statist. 33*(4), 1497–1537.

Beck, A. and M. Teboulle (2009, 01). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences 2*, 183–202.

Geer, S. v. d. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

Kilmer, M. and C. Martin (2011, 08). Factorization strategies for third-order tensors. *Linear Algebra Appl. 435*, 641–658.

Kilmer, M. E., C. D. Martin, and L. Perrone (2008). A third-order generaliza-

tion of the matrix svd as a product of third-order tensors. *Tufts University, Department of Computer Science, Tech. Rep. TR-2008-4*.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. New York, NY, USA: Springer.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat. 39*(5), 2302 – 2329.

Lu, C., J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan (2018, 04). Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Trans. Pattern Anal. Mach. Intell. 42*(4), 925–938.

Martin, C. D., R. Shafer, and B. LaRue (2013). An order-p tensor factorization with applications in imaging. *SIAM J. Sci. Comput. 35*(1), A474–A490.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York, NY, USA: Springer-Verlag.

Qin, W., H. Wang, F. Zhang, J. Wang, X. Luo, and T. Huang (2022). Low-rank high-order tensor completion with applications in visual data. *IEEE Trans. Image Process. 31*, 2433–2448.

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank so-

lutions of linear matrix equations via nuclear norm minimization. *SIAM REV. 52*(3), 471–501.

Song, Z., Y. Chen, and Z. Weihua (2025). Adaptively robust high-order tensor factorization for low-rank tensor reconstruction. *Pattern Recognit. 165*, 111600.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.