ISLET: Fast and Optimal Low-rank Tensor Regression via Importance Sketching

Anru Zhang[1], Yuetian Luo[1], Garvesh Raskutti[1], and Ming Yuan[2]

Abstract

In this paper, we develop a novel procedure for low-rank tensor regression, namely *Importance Sketching Low-rank Estimation for Tensors* (ISLET). The central idea behind ISLET is *importance sketching*, i.e., carefully designed sketches based on both the responses and low-dimensional structure of the parameter of interest. We show that the proposed method is sharply minimax optimal in terms of the mean-squared error under low-rank Tucker assumptions and under randomized Gaussian ensemble design. In addition, if a tensor is low-rank with group sparsity, our procedure also achieves minimax optimality. Further, we show through numerical studies that ISLET achieves comparable or better mean-squared error performance to existing state-of-the-art methods whilst having substantial storage and run-time advantages including capabilities for parallel and distributed computing. In particular, our procedure performs reliable estimation with tensors of dimension $p = O(10^8)$ and is 1 or 2 orders of magnitude faster than baseline methods.

Key words: dimension reduction, high-order orthogonal iteration, minimax optimality, sketching, tensor regression.

1 Introduction

The past decades have seen a large body of work on tenors or multiway arrays [65, 107, 32, 71]. Tensors arise in numerous applications involving multiway data (e.g., brain imaging [143], hyperspectral imaging [76], recommender system design [11]). In addition, tensor methods have been applied to many problems in statistics and machine learning where the observations are not necessarily tensors, such as topic and latent variable models [2], additive index models [5], high-order interaction pursuit [55], among others. In many of

[1] Department of Statistics, University of Wisconsin-Madison. (anruzhang@ stat.wisc.edu, yluo86@wisc.edu, raskutti@stat.wisc.edu)

[2] Department of Statistics, Columbia University (my2550@columbia.edu)

these settings, the tensor of interest is *high-dimensional* in that the ambient dimension, i.e, the dimension of the target parameter is substantially larger than the sample size. However in practice, the tensor parameter often has intrinsic dimension-reduced structure, such as low-rankness and sparsity [65, 112, 121], which makes inference possible. How to exploit such structure for tensors poses new *statistical* and *computational challenges* [103].

From a statistical perspective, a key question is how many samples are required to learn the suitable dimension-reduced structure and what the optimal mean-squared error rates are. Prior work has developed various tensor-based methods with theoretical guarantees based on regularization approaches [73, 91, 103, 117], the spectral method and projected gradient descent [29], alternating gradient descent [75, 113, 143], stochastic gradient descent [47], and power iteration methods [2]. However a number of these methods are not statistically optimal. Furthermore, some of these methods rely on evaluation of a full gradient, which is typically costly in the high-dimensional setting. This leads to computational challenges including both the *storage* of tensors and *run-time* of the algorithm.

From a computational perspective, one approach to address both the storage and run-time challenge is *randomized sketching.* Sketching methods have been widely studied (see e.g. [3, 4, 8, 14, 33, 34, 35, 37, 38, 56, 82, 92, 97, 99, 100, 102, 110, 111, 114, 118, 125, 126]). Many of these prior works on matrix or tensor sketching mainly focused on relative approximation error [14, 34, 92, 102] after randomized sketching which either may not yield optimal mean-squared error rates under statistical settings [102] or requires multiple sketching iterations [100, 101].

In this article, we address both computational and statistical challenges by developing a novel sketching-based estimating procedure for tensor regression. The proposed procedure is provably fast and sharply minimax optimal in terms of mean-squared error under randomized Gaussian design. The central idea lies in constructing specifically designed structural sketches, namely *importance sketching.* In contrast with randomized sketching methods, importance sketching utilizes both the response and structure of the target tensor parameter and reduces the dimension of parameters (i.e., the number of columns) instead of samples (i.e., the number of rows), which leads to statistical optimality whilst maintaining the computational advantages of many randomized sketching methods. See more comparison between importance sketching in this work and sketching in prior literature in Section 1.3.

1.1 Problem Statement

Specifically, we focus on the following low-rank tensor regression model,

$$y_j = \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}} \rangle + \varepsilon_j, \quad j = 1, \ldots, n, \tag{1}$$

where y_j and ε_j are responses and observation noise, respectively; $\{\boldsymbol{\mathcal{X}}_j\}_{j=1}^n$ are tensor covariates with randomized design; $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is the order-d tensor with parameters aligned in d ways. Here $\langle\cdot,\cdot\rangle$ stands for the usual vectorized inner product. The goal is to recover $\boldsymbol{\mathcal{A}}$ based on observations $\{y_j, \boldsymbol{\mathcal{X}}_j\}_{j=1}^n$. In particular, when $d = 2$, this becomes a low-rank matrix regression problem, which has been widely studied in recent years [25, 68, 104]. The main focus of this paper is solving the underdetermined equation system, where the sample size n is much smaller than the number of coefficients $\prod_{i=1}^d p_i$. This is because many applications belong to this regime. In particular, in the real data example to be discussed later, one MRI image is 121-by-145-by-121, which includes 2,122,945 parameters. Typically we can collect far less number of MRI images in practice.

The general regression model (1) includes specific problem instances with different choices of design $\boldsymbol{\mathcal{X}}$. Examples include matrix/tensor regression with general random or deterministic design [29, 77, 103, 143], matrix trace regression [6, 25, 43, 45, 68, 104], and matrix sparse recovery [132]. Another example is *matrix/tensor recovery via rank-1 projections* [18, 30, 55], which arise by setting $\boldsymbol{\mathcal{X}}_j = \mathbf{u}_j \circ \mathbf{v}_j \circ \mathbf{w}_j$, where $\mathbf{u}_j, \mathbf{v}_j, \mathbf{w}_j$ are random vectors and "$\circ$" represents the outer product, which includes phase retrieval [16, 23] as a special case. The very popular matrix/tensor completion example [27, 78, 90, 127, 128, 134] arises by setting $\boldsymbol{\mathcal{X}}_j = \left(\mathbf{e}_{a_j} \circ \mathbf{e}_{b_j} \circ \mathbf{e}_{c_j}\right)$, where $\mathbf{e}_j$ is the j-th canonical vector and $\{a_j, b_j, c_j\}_{j=1}^n$ are randomly selected integers from $\{1,\ldots,p_1\}\times\{1,\ldots,p_2\}\times\{1,\ldots,p_3\}$. Specific applications of this low-rank tensor regression model includes neuroimaging analysis [52, 75, 143], longitudinal relational data analysis [58], 3D imaging processing [53], etc.

For convenience of presentation, we specialize the discussions on order-3 tensors later, while the results can be extended to the general order-d tensors. In the modern high-dimensional setting, a variety of matrix/tensor data satisfy intrinsic structural assumptions, such as low-rankness [121] or sparsity [143], which makes the accurate estimation of $\boldsymbol{\mathcal{A}}$ possible even if the sample size n is smaller than the number of coefficients in the target tensor $\boldsymbol{\mathcal{A}}$. We thus focus on the low Tucker rank (r_1, r_2, r_3) tensor $\boldsymbol{\mathcal{A}}$ with the following Tucker decomposition [120]:

$$\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!] := \boldsymbol{\mathcal{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \tag{2}$$

where $\boldsymbol{\mathcal{S}}$ is a r_1-by-r_2-by-r_3 core tensor and $\mathbf{U}_k$ is a p_k-by-r_k matrix with orthonormal columns for $k = 1, 2, 3$. The rigorous definition of Tucker rank of a tensor and more discussions on tensor algebra are postponed to Section 2.1. In addition, the canonical polyadic (CP) low-rank tensors have also been widely considered in recent literature [55, 56, 113, 143]. Since any CP-rank-r tensor $\boldsymbol{\mathcal{A}} = \sum_{i=1}^r \lambda_i \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$ has the Tucker decomposition: $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{L}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$, where $\boldsymbol{\mathcal{L}}$ is the r-by-r-by-r diagonal tensor with diagonal entries $\lambda_1, \ldots, \lambda_r$, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$, and likewise for $\mathbf{B}, \mathbf{C}$ [65], our results naturally adapt to low CP-rank tensor regression. Also, with a slight abuse of notation, we will refer to low-rank and low Tucker rank interchangeably throughout the paper.

Moreover, we also consider a sparse setting where there may exist a subset of modes, say $J_s \subseteq \{1,2,3\}$, such that $\boldsymbol{\mathcal{A}}$ is sparse along these modes, i.e.

$$\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!], \quad \|\mathbf{U}_k\|_0 = \sum_{i=1}^{p_k} 1_{\{(\mathbf{U}_k)_{[i,:]} \neq 0\}} \leq s_k, \quad k \in J_s. \tag{3}$$

1.2 Our Contributions

We make the following major contributions to low-rank tensor regression in this article. Firstly, we introduce the main algorithm – *Importance Sketching Low-rank Estimation for Tensors* (ISLET). Our algorithm has three steps: (i) first we use the tensor technique high-order orthogonal iteration (HOOI) [36] or sparse tensor alternating thresholding - singular value decomposition (STAT-SVD) [136] to determine the importance sketching directions. Here, HOOI and STAT-SVD are regular and sparse tensor low-rank decomposition methods respectively, whose explanations are postponed to the forthcoming Sections 2.2 and 2.3; (ii) using the sketching directions from the first step, we perform importance sketching, then evaluate the dimension-reduced regression using the sketched tensors/matrices (to incorporate sparsity, we add a group-sparsity regularizer); (iii) we construct the final tensor estimator using the sketched components. Although the focus of this work is on low-rank tensor regression, we point out that our three-step procedure applies to general high-dimensional statistics problems with low-dimensional structure, provided that we can find a suitable projection operator in step (i), and inverse projection operator in step (iii).

One of the main advantages of ISLET is the scalability of the algorithm. The proposed procedure is computationally efficient due to the dimension reduction by importance sketchings. Most importantly, ISLET only require access to the full data twice, which significantly saves run time for large-scale settings when it is not possible to store all samples into the core memory. We also show that our algorithm can be naturally distributed across multiple machines that can significantly reduce computation time.

Secondly, we prove a deterministic oracle inequality for the ISLET procedure under the low-Tucker-rank assumption and general noise and design (Theorems 2 and 3). We additionally show that ISLET achieves the optimal mean-squared error (with the optimal constant for non-sparse ISLET) under randomized Gaussian design (Theorems 4, 5, 6, and 7). The following informal statement summarizes two of the main results of the article:

Theorem 1 (ISLET for tensor regression: informal)**.** *Consider the regular tensor regression problem with Gaussian ensemble design, where $\boldsymbol{\mathcal{A}}$ is Tucker rank-(r_1, r_2, r_3), $\boldsymbol{\mathcal{X}}_j$ has i.i.d. standard normal entries, $\varepsilon_j \overset{iid}{\sim} N(0, \sigma^2)$, and $\varepsilon_j, \boldsymbol{\mathcal{X}}_j$ are independent.*

(a) *Under regularity conditions, ISLET achieves the following optimal rate of convergence with the matching constant,*

$$\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^{2}=(1+o(1))\,\frac{m\sigma^{2}}{n},$$

where $m=r_1r_2r_3+r_1(p_1-r_1)+r_2(p_2-r_2)+r_3(p_3-r_3)$ *is exactly the degree of freedom of all Tucker rank-*(r_1,r_2,r_3) *tensors in* $\mathbb{R}^{p_1\times p_2\times p_3}$ *and* $\|\cdot\|_{\mathrm{HS}}$ *is the Hilbert-Schmidt norm to be defined in Section 2.1.*

(b) *If in addition,* (3) *holds with sparsity level* s_k*, then under regularity conditions, ISLET achieves the following optimal rate of convergence,*

$$\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^{2}\asymp\frac{m_s\sigma^{2}}{n},$$

where $m_s=r_1r_2r_3+\sum_{k\in J_s}s_k\left(r_k+\log(p_k/s_k)\right)+\sum_{k\notin J_s}p_kr_k$ *and "*$\asymp$*" denotes the asymptotic equivalence between two number series (see a more formal definition in Section 2.1).*

To the best of our knowledge, we are the first to develop the matching-constant optimal rate results for regular tensor regression under randomized Gaussian ensemble design, even for the low-rank matrix recovery case since it is not clear whether prior approaches (e.g. nuclear norm minimization) achieve sharp constants. We are also the first to develop the optimal rate results for tensor regression with sparsity condition (3).

Thirdly, proving the optimal mean-squared error bound presents a number of technical challenges and we introduce novel proof ideas to overcome these difficulties. In particular, one major difficulty lies in the analysis of reduced-dimensional regressions (see (7) in the forthcoming Section 2) since we analyze sketched regression models. To this end, we introduce partial linear models for these reduced-dimensional regressions from which we develop estimation error upper bounds.

The final and most important computational contribution is to display through numerical studies the advantages of our ISLET algorithms. Compared to state-of-the-art tensor estimation algorithms including non-convex projected gradient descent (PGD) [29], Tucker regression [143], and convex regularization [116], we show that our ISLET algorithm achieves comparable statistical performance with substantially faster computation. In particular, the runtime is 1-3 orders of magnitude faster than existing methods. In the most prominent example, our ISLET procedure can efficiently solve the ultrahigh-dimensional tensor regression with covariates of 7.68 terabytes. For the order-2 case, i.e., low-rank matrix regression, our simulation studies show that ISLET outperforms the classic nuclear norm minimization estimator. We also provide a real data application where we study the association between the attention-deficit/hyperactivity disorder disease and the high-dimensional MRI image tensors. We show that the proposed procedure provides significantly better prediction performance in much less time compared to state-of-the-art methods.

1.3 Related Literature

Our work is related to a broad range of literature varying from a number of communities including scientific computing, computer science, signal processing, applied mathematics, and statistics. Here we make an attempt to discuss existing results from these various communities however we do not claim that our literature survey is exhaustive.

Large-scale linear systems where the solution admits a low-rank tensor structure commonly arise after discretizing high-dimensional partial differential equations [59, 60, 80] and various methods have been proposed. For example, [12] developed algebraic and Gauss-Newton methods to solve the linear system with a CP low-rank tensor solution. [7, 10] proposed iterative projection methods to solve large-scale linear systems with Kronecker-product-type design matrices. [48] introduced a greedy approach. [69, 70] considered Riemannian optimization methods and tensor Krylov subspace methods, respectively. The readers are referred to [51] for a recent survey. Different from these works, our proposed ISLET is a one-step procedure that only involves solving a simple least squares regression after performing dimension reduction on covariates by importance sketching (see Steps 1 and 2 in Section 2.2). Moreover, many prior works mainly focused on computational aspects of their proposed methods [7, 13, 42, 48, 51], while we show that ISLET is not only computationally efficient (see more discussion and comparison on computation complexity in Section 2.2 Computation and Implementation part) but also has optimal theoretical guarantees in terms of mean square error under the statistical setting.

In addition, sketching methods play an important role in computation acceleration and has been widely considered in previous literature. For example, [34, 89, 92] provided accurate approximation algorithms based on sketching with novel embedding matrices, where the runtime is proportional to the number of the non-zero entries of the input matrix. Sketching methods have also been studied in robust ℓ_1 low-rank matrix approximation [85, 86, 88, 110, 141], general ℓ_p low-rank matrix approximation [8, 31], low-rank tensor approximation [111], etc. In the regression context, the sketching method has been considered for the least squares regression [34, 37, 92, 101, 102], ℓ_p regression [34, 89, 92], Kronecker product regression [37], ridge regression [3, 124], regularized kernel regression [22, 140], etc. Various types of random sketching matrices have been developed, including random sub-Gaussian [101], random sampling [39, 40], CountSketch [28, 33], Sparse Johnson-Lindenstrauss transformation [64], among many others. The readers are also referred to survey papers on sketching by Mahoney [82] and Woodruff [126]. The proposed method in this paper is different from these previous works in various aspects. First, many randomized sketching methods in the literature focused on relative approximation error [82, 126] and the sketching matrices are constructed only based on covariates [39, 40, 64, 101, 102]. In contrast, we explicitly construct "supervised" sketching matrices based on both the response y_j and covariates $\boldsymbol{\mathcal{X}}_j$ and obtain optimal bounds in mean square error under the statistical setting. Second, essentially speaking, our proposed importance sketching scheme reduces the number of columns (parameters) instead of the number of rows (samples) in the linear equation system. Third, different from the sketching on an overdetermined system of least squares [34, 37, 92, 101, 102], we mainly focus on the high-dimensional setting where the number of samples can be significantly smaller than the number of coefficients.

1.4 Organization

In Section 2.1 we introduce important notation; then we present our ISLET procedure under non-sparse and sparse settings in Sections 2.2 and 2.3, respectively and illustrate the procedure from a sketching perspective in Section 2.4; in Section 3 we provide general theoretical guarantees for our procedure which make no assumptions on the design or the noise distribution; in Section 4 we specialize our bounds to tensor regression with low Tucker rank and assume the design is independent Gaussian; a simulation study showing the substantial computational benefits of our algorithm are provided in Section 5. Additional notation, discussion on general-order ISLET, simulation results, an application to attention deficit hyperactivity disorder (ADHD) MRI Imaging data analysis, and all technical proofs are provided in the supplementary materials [137].

2 Our Procedure: ISLET

In this section, we introduce the general procedure of importance sketching low-rank estimation for tensors (ISLET). Although for ease of presentation we will focus on order-3 tensors, the procedure for the general order-d case can also be treated. Details of matrices and tensors greater than order 3 are provided in Section C of the supplementary material [137].

2.1 Notation and Preliminaries

The following notation will be used throughout this article. Additional definitions can be found in Section A in the supplementary materials. Lowercase letters (e.g., a, b), lowercase boldface letters (e.g. $\mathbf{u}, \mathbf{v}$), uppercase boldface letters (e.g., $\mathbf{U}, \mathbf{V}$), and boldface calligraphic letters (e.g., $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}$) are used to denote scalars, vectors, matrices, and order-3-or-higher tensors respectively. For simplicity, we denote $\boldsymbol{\mathcal{X}}_j$ as the tensor indexed by j in a sequence of tensors $\{\boldsymbol{\mathcal{X}}_j\}$. For any two series of numbers, say $\{a_i\}$ and $\{b_i\}$, denote $a \asymp b$ if there exist uniform constants $c, C > 0$ such that $ca_i \leq b_i \leq Ca_i, \forall i$ and $a = \Omega(b)$ if there exists uniform constant $c > 0$ such that $a_i \geq cb_i, \forall i$. We use bracket subscripts to denote sub-vectors, sub-matrices, and sub-tensors. For example, $\mathbf{v}_{[2:r]}$ is the vector with the 2nd to rth entries of $\mathbf{v}$; $\mathbf{D}_{[i_1,i_2]}$ is the entry of $\mathbf{D}$ on the i_1-th row and i_2-th column; $\mathbf{D}_{[(r+1):p_1,:]}$ contains the $(r+1)$-th to the p_1-th rows of $\mathbf{D}$; $\boldsymbol{\mathcal{A}}_{[1:s_1,1:s_2,1:s_3]}$ is the s_1-by-s_2-by-s_3 sub-tensor of $\boldsymbol{\mathcal{A}}$ with index set $\{(i_1,i_2,i_3) : 1 \leq i_1 \leq s_1, 1 \leq i_2 \leq s_2, 1 \leq i_3 \leq s_3\}$. For any vector $\mathbf{v} \in \mathbb{R}^{p_1}$, define its ℓ_q norm as $\|\mathbf{v}\|_q = (\sum_i |v_i|^q)^{1/q}$. For any matrix $\mathbf{D} \in \mathbb{R}^{p_1 \times p_2}$, let $\sigma_k(\mathbf{D})$ be the k-th singular value of $\mathbf{D}$. In particular, the least non-trivial singular value of $\mathbf{D}$, defined as $\sigma_{\min}(\mathbf{D}) = \sigma_{p_1 \wedge p_2}(\mathbf{D})$, will be extensively used in later analysis. We also denote $\mathrm{SVD}_r(\mathbf{D}) = [\mathbf{u}_1 \ \cdots \mathbf{u}_r]$ and $\mathrm{QR}(\mathbf{D})$ as the subspace composed of the leading r left singular vectors and the Q part of the QR orthogonalization of $\mathbf{D}$, respectively. The matrix Frobenius and spectral norms are defined as $\|\mathbf{D}\|_F = \left(\sum_{i_1,i_2} \mathbf{D}^2_{[i_1,i_2]}\right)^{1/2} = (\sum_{i=1}^{p_1 \wedge p_2} \sigma_i^2(\mathbf{D}))^{1/2}$ and $\|\mathbf{D}\| = \max_{\mathbf{u} \in \mathbb{R}^{p_2}} \|\mathbf{D}\mathbf{u}\|_2 / \|\mathbf{u}\|_2 = \sigma_1(\mathbf{D})$. In addition, $\mathbf{I}_r$ represents the r-by-r identity matrix. Let $\mathbb{O}_{p,r} = \{\mathbf{U} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r\}$ be the set of all p-by-r matrices with orthonormal columns. For any $\mathbf{U} \in \mathbb{O}_{p,r}$, $P_{\mathbf{U}} = \mathbf{U}\mathbf{U}^\top$ represents the projection matrix onto the column space of $\mathbf{U}$; we also use $\mathbf{U}_\perp \in \mathbb{O}_{p,p-r}$ to represent the orthonormal complement of $\mathbf{U}$. For any event A, let $\mathbb{P}(A)$ be the probability that A occurs.

For any matrix $\mathbf{D} \in \mathbb{R}^{p_1 \times p_2}$ and order-d tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, let $\mathrm{vec}(\mathbf{D})$ and $\mathrm{vec}(\boldsymbol{\mathcal{A}})$ be the vectorization of $\mathbf{D}$ and $\boldsymbol{\mathcal{A}}$, respectively. The matricization $\mathcal{M}(\cdot)$ is the operation that unfolds or flattens the order-d tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ into the matrix $\mathcal{M}_k(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{p_k \times \prod_{j \neq k} p_j}$ for $k = 1, \ldots, d$. Since the formal entry-wise definitions of matricization and vectorization is rather tedious, we leave them to Section A in the supplementary materials [137]. The Hilbert-Schmidt norm is defined as $\|\boldsymbol{\mathcal{A}}\|_{\mathrm{HS}} = \left(\sum_{i_1,\ldots,i_d} \boldsymbol{\mathcal{A}}^2_{[i_1,\ldots,i_d]}\right)^{1/2}$. An order-$d$ tensor is rank-one if it can be written as the outer product of d nonzero vectors. The CP-rank of any tensor $\boldsymbol{\mathcal{A}}$ is defined as the minimal number r such that $\boldsymbol{\mathcal{A}}$ can be decomposed as $\boldsymbol{\mathcal{A}} = \sum_{i=1}^r \boldsymbol{\mathcal{B}}_i$ for rank-1 tensors $\boldsymbol{\mathcal{B}}_i$. The Tucker rank (or multilinear rank) of a tensor $\boldsymbol{\mathcal{A}}$ is defined as a d-tuple $(r_1, \ldots, r_d)$, where $r_k = \mathrm{rank}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))$. The k-mode product of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \ldots \times p_d}$ with a matrix $\mathbf{U} \in \mathbb{R}^{p_k \times r_k}$ is denoted by $\boldsymbol{\mathcal{A}} \times_k \mathbf{U}$ and is of size $p_1 \times \cdots \times p_{k-1} \times r_k \times p_{k+1} \times \cdots \times p_d$, such that

$$(\boldsymbol{\mathcal{A}} \times_k \mathbf{U})_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_d]} = \sum_{i_k=1}^{p_k} \boldsymbol{\mathcal{A}}_{[i_1,i_2,\ldots,i_d]} \mathbf{U}_{[i_k,j]}.$$

For convenience of presentation, all mode indices $(\cdot)_k$ of an order-3 tensor are in the sense of modulo-3, e.g., $r_1 = r_4$, $s_2 = s_5$, $p_0 = p_3$, $\boldsymbol{\mathcal{X}} \times_4 \mathbf{U}_4 = \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}_1$.

For any matrices $\mathbf{U} \in \mathbb{R}^{p_1 \times p_2}$ and $\mathbf{V} \in \mathbb{R}^{m_1 \times m_2}$, let

$$\mathbf{U} \otimes \mathbf{V} = \begin{bmatrix} \mathbf{U}_{[1,1]} \cdot \mathbf{V} & \cdots & \mathbf{U}_{[1,p_2]} \cdot \mathbf{V} \\ \vdots & & \vdots \\ \mathbf{U}_{[p_1,1]} \cdot \mathbf{V} & \cdots & \mathbf{U}_{[p_1,p_2]} \cdot \mathbf{V} \end{bmatrix} \in \mathbb{R}^{(p_1 m_1) \times (p_2 m_2)}$$

be the Kronecker product. Some intrinsic identities among Kronecker product, vectorization, and matricization, which will be used later in this paper, are summarized in Lemma 1 in the supplementary materials [137]. The readers can refer to [65] for a more comprehensive introduction to tensor algebra. Finally, we use C, C_1, C_2, c and other variations to represent the large and small constants, whose actual value may vary from line to line.

2.2 Regular Low-rank Tensor Recovery

We first consider the tensor regression model (1), where $\boldsymbol{\mathcal{A}}$ is low-rank (2) without sparsity assumptions. The proposed algorithm of ISLET is divided into three steps and a pictorial illustration is provided in Figures 1 - 3 for readers' better understanding. The pseudo-code is provided in Algorithm 1.

Step 1 (Probing importance sketching directions) We first probe the importance sketching directions. When the covariates satisfy $\mathbb{E}\text{vec}(\boldsymbol{\mathcal{X}}_j)\text{vec}(\boldsymbol{\mathcal{X}}_j)^\top = \mathbf{I}_{p_1p_2p_3}$, we evaluate

$$\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{j=1}^{n} y_j \boldsymbol{\mathcal{X}}_j. \tag{4}$$

$\widetilde{\boldsymbol{\mathcal{A}}}$ is essentially the covariance tensor between y and $\boldsymbol{\mathcal{X}}$. Since $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!]$ has low Tucker rank, we perform the high-order orthogonal iterations (HOOI) on $\widetilde{\boldsymbol{\mathcal{A}}}$ to obtain $\widetilde{\mathbf{U}}_k \in \mathbb{O}_{p_k, r_k}, k = 1, 2, 3$ as initial estimates for $\mathbf{U}_k$. Here, HOOI is a classic method for tensor decomposition that can be traced back to Lathauwer, Moor, and Vandewalle [36]. The central idea of HOOI is the power iterated singular value thresholding. Then, the outcome of HOOI $\{\widetilde{\mathbf{U}}_k\}_{k=1}^3$ yield the following low-rank approximation for $\boldsymbol{\mathcal{A}}$,

$$\boldsymbol{\mathcal{A}} \approx [\![\widetilde{\boldsymbol{\mathcal{S}}}; \widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3]\!], \quad \text{where} \quad \widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!] \in \mathbb{R}^{r_1 \times r_2 \times r_3}. \tag{5}$$

We further evaluate

$$\widetilde{\mathbf{V}}_k := \text{QR}\left(\mathcal{M}_k^\top(\widetilde{\boldsymbol{\mathcal{S}}})\right) \in \mathbb{O}_{r_{k+1}r_{k+2}, r_k}, \quad k = 1, 2, 3.$$

$\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k\}_{k=1}^3$ obtained here are regarded as the *importance sketching directions.* As we will further illustrate in Section 3.1, the combinations of $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{V}}_k$ provide approximations for singular subspaces of $\mathcal{M}_k(\boldsymbol{\mathcal{A}})$.

Step 2 (Linear regression on sketched covariates) Next, we perform sketching to reduce the dimension of the original regression model (1). To be specific, we project the original high-dimensional covariates onto the dimension-reduced subspace "that is important in the covariance between y and $\boldsymbol{\mathcal{X}}$" and construct the following *importance sketching covariates*,

$$\begin{aligned}
&\widetilde{\mathbf{X}} = \left[\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_1} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_2} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_3}\right] \in \mathbb{R}^{n\times m},\\
&\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n\times m_{\boldsymbol{\mathcal{B}}}}, \quad \left(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\right)_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top\right),\\
&\widetilde{\mathbf{X}}_{\mathbf{D}_k} \in \mathbb{R}^{n\times m_{\mathbf{D}_k}}, \quad \left(\widetilde{\mathbf{X}}_{\mathbf{D}_k}\right)_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right) \widetilde{\mathbf{V}}_k\right),
\end{aligned} \tag{6}$$

where $m_{\boldsymbol{\mathcal{B}}} = r_1 r_2 r_3$, $m_{\mathbf{D}_k} = (p_k - r_k) r_k$, $k = 1, 2, 3$, and $m = m_{\boldsymbol{\mathcal{B}}} + m_{\mathbf{D}_1} + m_{\mathbf{D}_2} + m_{\mathbf{D}_3}$. Then, we evaluate the least-squares estimator of the sub-model with importance sketching covariates $\widetilde{\mathbf{X}}$,

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^m}{\arg\min} \left\| y - \widetilde{\mathbf{X}}\boldsymbol{\gamma} \right\|_2^2. \tag{7}$$

The dimension of sketching covariate regression (7) is m, which is significantly smaller than the dimension of the original tensor regression model, $p_1 p_2 p_3$. Consequently, the computational cost can be significantly reduced.

Step 3 (Assembling the final estimate) Then, $\widehat{\boldsymbol{\gamma}}$ is divided into four segments according to the block-wise structure of $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}, \widetilde{\mathbf{X}}_{\mathbf{D}_1}, \widetilde{\mathbf{X}}_{\mathbf{D}_2}, \widetilde{\mathbf{X}}_{\mathbf{D}_3}]$,

$$\begin{aligned}
\mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) &= \widehat{\boldsymbol{\gamma}}_{[1:m_{\boldsymbol{\mathcal{B}}}]},\\
\mathrm{vec}(\widehat{\mathbf{D}}_1) &= \widehat{\boldsymbol{\gamma}}_{[(m_{\boldsymbol{\mathcal{B}}}+1):(m_{\boldsymbol{\mathcal{B}}}+m_{\mathbf{D}_1})]},\\
\mathrm{vec}(\widehat{\mathbf{D}}_2) &= \widehat{\boldsymbol{\gamma}}_{[(m_{\boldsymbol{\mathcal{B}}}+m_{\mathbf{D}_1}+1):(m_{\boldsymbol{\mathcal{B}}}+m_{\mathbf{D}_1}+m_{\mathbf{D}_2})]},\\
\mathrm{vec}(\widehat{\mathbf{D}}_3) &= \widehat{\boldsymbol{\gamma}}_{[(m_{\boldsymbol{\mathcal{B}}}+m_{\mathbf{D}_1}+m_{\mathbf{D}_2}+1):(m_{\boldsymbol{\mathcal{B}}}+m_{\mathbf{D}_1}+m_{\mathbf{D}_2}+m_{\mathbf{D}_3})]}.
\end{aligned} \tag{8}$$

Finally, we construct the regression estimator $\widehat{\boldsymbol{\mathcal{A}}}$ for the original problem (1) using the regression estimator $\widehat{\boldsymbol{\gamma}}$ for the sub-model (8): let $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$ and calculate

$$\widehat{\mathbf{L}}_k = \left(\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widehat{\mathbf{D}}_k\right)\left(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right)^{-1}, \quad k = 1, 2, 3, \quad \widehat{\boldsymbol{\mathcal{A}}} = \left[\!\left[\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{L}}_1, \widehat{\mathbf{L}}_2, \widehat{\mathbf{L}}_3\right]\!\right]. \tag{9}$$

More interpretation of (9) is given in Section 3.1.

Remark 1 (Alternative Construction of $\widetilde{\boldsymbol{\mathcal{A}}}$ in Step 1)**.** *When* $\mathbb{E}\mathrm{vec}(\boldsymbol{\mathcal{X}})\mathrm{vec}(\boldsymbol{\mathcal{X}})^\top \neq \mathbf{I}_{p_1 p_2 p_3}$*, we could consider the following alternative ways to construct the initial estimate* $\widetilde{\boldsymbol{\mathcal{A}}}$*. Firstly, in some cases we could do construction depending on the covariance structure of* $\boldsymbol{\mathcal{X}}$*. For example, in the framework of tensor recovery via rank-one sketching (discussed in the introduction), we have* $\boldsymbol{\mathcal{X}}_j = \mathbf{u}_j \circ \mathbf{u}_j \circ \mathbf{u}_j$ *and* $\mathbf{u}_j \in \mathbb{R}^p$ *has iid entry* $N(0,1)$*. By the high-order*

Stein's identity [63], one can show that

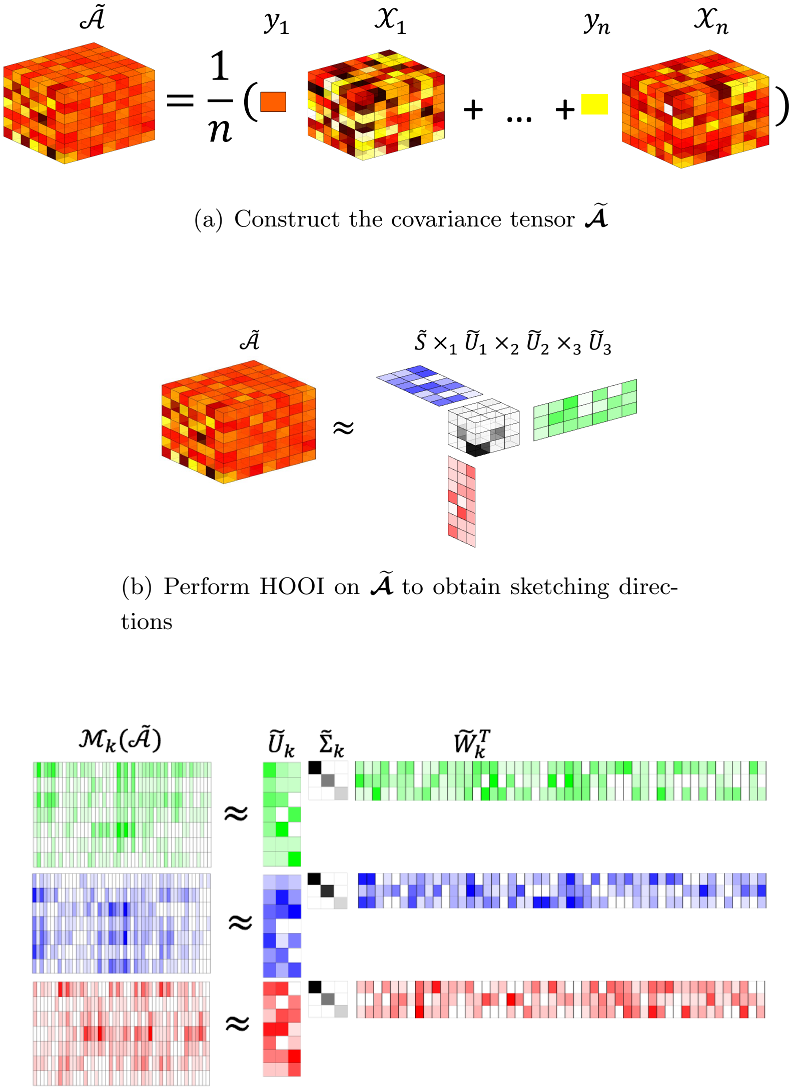
$$\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{6}\left[\frac{1}{n}\sum_{j=1}^{n} y_j \mathbf{u}_j \circ \mathbf{u}_j \circ \mathbf{u}_j - \sum_{j=1}^{p} \left(\mathbf{w}\circ\mathbf{e}_j\circ\mathbf{e}_j + \mathbf{e}_j\circ\mathbf{w}\circ\mathbf{e}_j + \mathbf{e}_j\circ\mathbf{e}_j\circ\mathbf{w}\right)\right],$$

is a proper initial unbiased estimator for $\boldsymbol{\mathcal{A}}$ *[55, Lemma 4]. Here,* $\mathbf{w} = \frac{1}{n}\sum_{i=1}^{n} y_j \mathbf{u}_j$, $\mathbf{e}_j$ *is the* j*th canonical basis in* $\mathbb{R}^p$. *Another commonly used setting in data analysis is the high-order Kronecker covariance structure:* $\mathbb{E}(\mathrm{vec}(\boldsymbol{\mathcal{X}}_j)\mathrm{vec}(\boldsymbol{\mathcal{X}}_j)^\top) = \boldsymbol{\Sigma}_3 \otimes \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, *where* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p_k \times p_k}, k = 1, 2, 3$ *are covariance matrices along three modes, respectively [57, 81, 84, 98, 144]. Under this assumption, we can first apply existing approaches to obtain estimators* $\widehat{\boldsymbol{\Sigma}}_k$ *for* $\boldsymbol{\Sigma}_k$, *then whiten the covariates by replacing* $\boldsymbol{\mathcal{X}}_j$ *by* $[\![\boldsymbol{\mathcal{X}}_j; \widehat{\boldsymbol{\Sigma}}_1^{-1/2}, \widehat{\boldsymbol{\Sigma}}_2^{-1/2}, \widehat{\boldsymbol{\Sigma}}_3^{-1/2}]\!]$. *After this pre-processing step, the other steps of ISLET still follow. Moreover, it still remains an open question how to perform initialization if* $\boldsymbol{\mathcal{X}}$ *has the more general, unstructured, and unknown design.*

Remark 2 (Alternative Methods to HOOI)**.** *In addition to high-order orthogonal iteration (HOOI), there are a variety of methods proposed in the literature to compute the low-rank tensor approximation, such as Newton-type optimization methods on manifolds [41, 61, 62, 106], black box approximation [9, 21, 83, 94, 95, 135], generalizations of Krylov subspace method [49, 105], greedy approximation method [48], among many others. Further, black box approximation methods [9, 21, 94, 95, 135] can be applied even if the initial estimator* $\widetilde{\boldsymbol{\mathcal{A}}}$ *does not fit into the core memory. When the tensor is further approximately CP low-rank, we can also apply the randomized compressing method [108, 109] or randomized block sampling [123] to obtain the CP low-rank tensor approximation. Although the rest of our discussion will focus on the HOOI procedure for initialization, these alternative methods can also be applied to obtain an initialization for the ISLET algorithm.*

Computation and implementation. We briefly discuss computational complexity and implementation aspects for the ISLET procedure here. It is noteworthy that ISLET accesses the sample only twice for constructing the covariance tensor (Step 1) and importance sketching covariates (Step 2), respectively. In large scale cases where it is difficult to store the whole dataset into random-access memory (RAM), this advantage can highly save the computational costs.

In addition, in the order-3 tensor case, when each mode shares the same dimension $p_k = p$ and rank $r_k = r$, the total number of observable values is $O(np^3)$ and the time complexity of ISLET is $O\left(np^3r + nr^6 + Tp^4\right)$ where T is the number of HOOI iterations. For general order-d tensor regression, time complexity of ISLET is $O\left(np^dr + nr^{2d} + Tp^{d+1}\right)$. In contrast, the time complexity of the non-convex PGD [29] is $O\left(T'(np^d + rp^{d+1})\right)$, where T' is the number of iterations of gradient descent; [13] introduced an optimization based

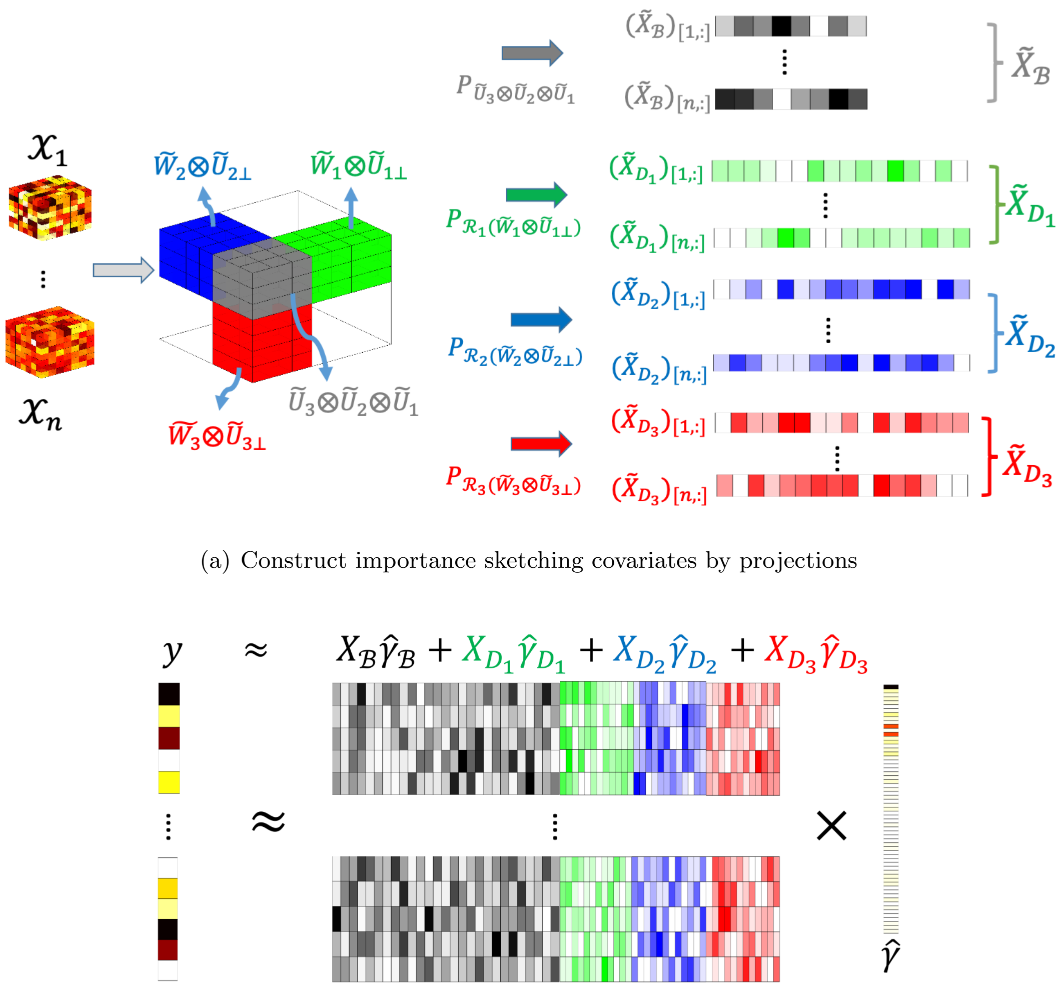


(a) Construct the covariance tensor $\widetilde{\boldsymbol{\mathcal{A}}}$

(b) Perform HOOI on $\widetilde{\boldsymbol{\mathcal{A}}}$ to obtain sketching directions

(c) The sketching directions yield low-rank approximations for $\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{A}}})$

Figure 1: Illustration for Step 1 of ISLET


$\mathcal{X}_1$
$\mathcal{X}_n$
$\widetilde{W}_2\otimes\widetilde{U}_{2\perp}$
$\widetilde{W}_1\otimes\widetilde{U}_{1\perp}$
$\widetilde{W}_3\otimes\widetilde{U}_{3\perp}$
$\widetilde{U}_3\otimes\widetilde{U}_2\otimes\widetilde{U}_1$
$P_{\widetilde{U}_3\otimes\widetilde{U}_2\otimes\widetilde{U}_1}$
$(\widetilde{X}_{\mathcal{B}})_{[1,:]}$
$(\widetilde{X}_{\mathcal{B}})_{[n,:]}$
$\widetilde{X}_{\mathcal{B}}$
$P_{\mathcal{R}_1(\widetilde{W}_1\otimes\widetilde{U}_{1\perp})}$
$(\widetilde{X}_{D_1})_{[1,:]}$
$(\widetilde{X}_{D_1})_{[n,:]}$
$\widetilde{X}_{D_1}$
$P_{\mathcal{R}_2(\widetilde{W}_2\otimes\widetilde{U}_{2\perp})}$
$(\widetilde{X}_{D_2})_{[1,:]}$
$(\widetilde{X}_{D_2})_{[n,:]}$
$\widetilde{X}_{D_2}$
$P_{\mathcal{R}_3(\widetilde{W}_3\otimes\widetilde{U}_{3\perp})}$
$(\widetilde{X}_{D_3})_{[1,:]}$
$(\widetilde{X}_{D_3})_{[n,:]}$
$\widetilde{X}_{D_3}$
$y \approx X_{\mathcal{B}}\hat{\gamma}_{\mathcal{B}} + X_{D_1}\hat{\gamma}_{D_1} + X_{D_2}\hat{\gamma}_{D_2} + X_{D_3}\hat{\gamma}_{D_3}$
$\approx$
$\times$
$\hat{\gamma}$

(a) Construct importance sketching covariates by projections

(b) Perform regression of submodel with importance sketching covariates

Figure 2: Illustration for Step 2 of ISLET

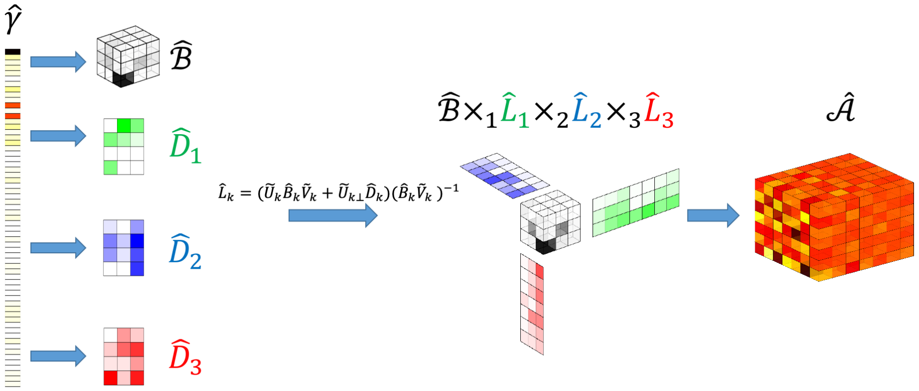

$\hat{\gamma}$
$\widehat{\mathcal{B}}$
$\widehat{D}_1$
$\widehat{D}_2$
$\widehat{D}_3$
$\hat{L}_k = (\widetilde{U}_k\hat{B}_k\widetilde{V}_k + \widetilde{U}_{k\perp}\hat{D}_k)(\hat{B}_k\widetilde{V}_k)^{-1}$
$\widehat{\mathcal{B}}\times_1\hat{L}_1\times_2\hat{L}_2\times_3\hat{L}_3$
$\hat{\mathcal{A}}$

Figure 3: Illustration for Step 3 of ISLET

method with time complexity $O(T'dnp^dr)$ where T' is the number of iterations in Gauss-Newton method. We can see if $T' \geq r$, a typical situation in practice, ISLET is significantly faster than these previous methods.

It is worth pointing out that the computing time of ISLET is still high when the tensor parameter has a large order d. In fact, without any structural assumption on the design tensors $\boldsymbol{\mathcal{X}}_j$, such a time cost may be unavoidable since reading in all data requires $O(np^d)$ operations. If there is extra structure on the design tensor, e.g., Kronecker product [7, 59, 60, 80] and low separation rank [10, 48], the computing time can be significantly reduced by applying methods in this body of literature. Here, we mainly focus on the setting where $\boldsymbol{\mathcal{X}}_j$ does not satisfy a clear structural assumption since in many real data applications, e.g., the neuroimaging data example studied in this and many other works [1, 77, 113, 143], the design tensors $\boldsymbol{\mathcal{X}}_j$ may not have a clear known structure.

Moreover, in the order-3 tensor case, instead of storing all $\{\boldsymbol{\mathcal{X}}_j\}_{j=1}^n$ in the memory which requires $O(np^3)$ RAM, ISLET only requires $O(p^3 + n(pr + r^3))$ RAM space if one chooses to access the samples from hard disks but not to store to RAM. This makes large-scale computing possible. We empirically investigate the computation cost by simulation studies in Section 5.

The proposed ISLET procedure also allows convenient parallel computing. Suppose we distribute all n samples across B machines: $\{(\boldsymbol{\mathcal{X}}_{bi}, y_{bi})\}_{i=1}^{B_b}$, $b = 1, \ldots, B$, where $B_b \approx n/B$. To evaluate the covariance tensor in Step 1, we can calculate $\widetilde{\boldsymbol{\mathcal{A}}}_b = \sum_{i=1}^{B_i} y_{bi}\boldsymbol{\mathcal{X}}_{bi}$ in each machine, then summarize them as $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{b=1}^{B} \widetilde{\boldsymbol{\mathcal{A}}}_b$; to construct sketching covariates and perform partial regression in Step 2, we calculate

$$\mathbf{y}_b = (y_{b1}, \ldots, y_{bB_b})^\top \in \mathbb{R}^{B_b}, \tag{10}$$

$$\begin{aligned}
\widetilde{\mathbf{X}}_{bi} &= \left[\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}},bi} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_1,bi} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_2,bi} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_3,bi}\right] \in \mathbb{R}^m, \\
\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}},bi} &= \operatorname{vec}\left(\boldsymbol{\mathcal{X}}_{bi} \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top\right), \qquad (11) \\
\widetilde{\mathbf{X}}_{\mathbf{D}_k,bi} &= \operatorname{vec}\left(\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_{bi} \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right) \widetilde{\mathbf{V}}_k\right),
\end{aligned}$$

$$\widetilde{\mathbf{G}}_b = \sum_{i=1}^{B_b} \widetilde{\mathbf{X}}_{bi}^\top \widetilde{\mathbf{X}}_{bi}, \quad \widetilde{\mathbf{z}}_b = \sum_{i=1}^{B_b} \widetilde{\mathbf{X}}_{bi}^\top y_{bi} \tag{12}$$

in each machine. Then we combine the outcomes to

$$\widehat{\gamma} = \left(\sum_{b=1}^{B} \widetilde{\mathbf{G}}_b\right)^{-1} \left(\sum_{b=1}^{B} \widetilde{\mathbf{z}}_b\right).$$

The computational complexity can be reduced to $O\left(\frac{np^3r + nr^6}{B} + Tp^4\right)$ via the parallel scheme. In the large-scale simulation we present in this article, we implement this parallel scheme for speed-up.

To implement the proposed procedure, the input of Tucker rank are required as tuning parameters. When they are unknown in practice, we can perform cross-validation or an adaptive rank selection scheme. A more detailed description and numerical results are postponed to Section D in the supplementary materials [137].

2.3 Sparse Low-rank Tensor Recovery

When the target tensor $\boldsymbol{\mathcal{A}}$ is simultaneously low-rank and sparse, in the sense that (3) holds for a subset $J_s \subseteq \{1,2,3\}$ known a priori, we introduce the following sparse ISLET procedure. The pseudo-code for sparse ISLET is summarized in Algorithm 2.

Step 1 (Probing sketching directions) When $\mathbb{E}\text{vec}(\boldsymbol{\mathcal{X}})\text{vec}(\boldsymbol{\mathcal{X}})^\top = \mathbf{I}_{p_1p_2p_3}$, we still evaluate the covariance tensor $\widetilde{\boldsymbol{\mathcal{A}}}$ as Equation (4). Noting that $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!]$ and $\{\mathbf{U}_k\}_{k\in J_s}$ are row-wise sparse, we apply the sparse tensor alternating thresholding SVD (STAT-SVD) [136] on $\widetilde{\boldsymbol{\mathcal{A}}}$ to obtain $\widetilde{\mathbf{U}}_k \in \mathbb{O}_{p_k,r_k}, k = 1,2,3$ as initial estimates for $\mathbf{U}_k$. Here, STAT-SVD is a sparse tensor decomposition method proposed by [136] with central ideas of the double projection & thresholding scheme and power iteration. Via STAT-SVD, we obtain the following sparse and low-rank approximation of $\boldsymbol{\mathcal{A}}$,

$$\boldsymbol{\mathcal{A}} \approx [\![\widetilde{\boldsymbol{\mathcal{S}}}; \widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3]\!], \quad \widetilde{\mathbf{U}}_k \in \mathbb{O}_{p_k,r_k}, \quad \widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!] \in \mathbb{R}^{r_1\times r_2\times r_3}.$$

We further evaluate

$$\widetilde{\mathbf{V}}_k = \text{QR}\left(\mathcal{M}_k^\top(\widetilde{\boldsymbol{\mathcal{S}}})\right) \in \mathbb{O}_{r_{k+1}r_{k+2},r_k}.$$

Step 2 (Group Lasso on sketched covariates) We perform sketching and construct the following importance sketching covariates based on $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k\}_{k=1}^3$,

$$\begin{aligned} &\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n\times(r_1r_2r_3)}, \quad (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[i,:]} = \text{vec}\left(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top\right), \\ &\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n\times p_kr_k}, \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[i,:]} = \text{vec}\left(\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right)\widetilde{\mathbf{V}}_k\right). \end{aligned} \tag{13}$$

Then we perform regression on sub-models with these reduced-dimensional covariates $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}$ and $\widetilde{\mathbf{X}}_{\mathbf{E}_k}$ respectively using least squares and group Lasso [46, 133],

$$\widehat{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{r_1\times r_2\times r_3}, \quad \text{vec}(\widehat{\boldsymbol{\mathcal{B}}}) = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{r_1r_2r_3}}{\arg\min} \|y - \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\boldsymbol{\gamma}\|_2^2, \tag{14}$$

$$\widehat{\mathbf{E}}_k \in \mathbb{R}^{p_k\times r_k}, \text{vec}(\widehat{\mathbf{E}}_k) = \begin{cases} \arg\min_{\boldsymbol{\gamma}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2, & \text{if } k \notin J_s; \\ \arg\min_{\boldsymbol{\gamma}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2 + \eta_k \sum_{j=1}^{p_k} \|\boldsymbol{\gamma}_{G_j^k}\|_2, & \text{if } k \in J_s. \end{cases} \tag{15}$$

Here, $\{\eta_k\}_{k\in J_s}$ are the penalization level and

$$G_j^k = \{j, j+p_k, \ldots, j+p_k(r_k-1)\}, \quad j = 1,\ldots,p_k \tag{16}$$

form a partition of $\{1, \ldots, p_k r_k\}$ that is induced by the construction of $\widetilde{\mathbf{X}}_{\mathbf{E}_k}$ (details for why using group lasso can be found in Section 3.2).

Step 3 (Constructing the final estimator) $\widehat{\boldsymbol{\mathcal{A}}}$ can be constructed using the regression coefficients $\widehat{\boldsymbol{\mathcal{B}}}$ and $\widehat{\mathbf{E}}_k$'s in the submodels (14) and (15),

$$\widehat{\boldsymbol{\mathcal{A}}} = \left[\!\!\left[\widehat{\boldsymbol{\mathcal{B}}}, (\widehat{\mathbf{E}}_1(\widetilde{\mathbf{U}}_1^\top \widehat{\mathbf{E}}_1)^{-1}), (\widehat{\mathbf{E}}_2(\widetilde{\mathbf{U}}_2^\top \widehat{\mathbf{E}}_2)^{-1}), (\widehat{\mathbf{E}}_3(\widetilde{\mathbf{U}}_3^\top \widehat{\mathbf{E}}_3)^{-1}) \right]\!\!\right]. \tag{17}$$

More interpretation of (17) can be found in Section 3.2.

Algorithm 1 Importance Sketching Low-rank Estimation for Tensors (ISLET): Order-3 Case

```
Input: sample $\{y_j, \boldsymbol{\mathcal{X}}_j\}_{j=1}^n$, Tucker rank $\boldsymbol{r} = (r_1, r_2, r_3)$.
Calculate $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{j=1}^n y_j \boldsymbol{\mathcal{X}}_j$.
Apply HOOI on $\widetilde{\boldsymbol{\mathcal{A}}}$ and obtain initial estimates $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3$.
Let $\widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!]$. Evaluate the sketching direction,
    $$\widetilde{\mathbf{V}}_k = \mathrm{QR}\left[\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right], \quad k = 1, 2, 3.$$
Construct $\widetilde{\mathbf{X}} = \left[\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\ \widetilde{\mathbf{X}}_{\mathbf{D}_1}\ \widetilde{\mathbf{X}}_{\mathbf{D}_2}\ \widetilde{\mathbf{X}}_{\mathbf{D}_3}\right] \in \mathbb{R}^{n\times m}$, where
    $$\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n\times m_{\boldsymbol{\mathcal{B}}}}, (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top\right),$$
    $$\widetilde{\mathbf{X}}_{\mathbf{D}_k} \in \mathbb{R}^{n\times m_{\mathbf{D}_k}}, (\widetilde{\mathbf{X}}_{\mathbf{D}_k})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right) \widetilde{\mathbf{V}}_k\right),$$
    for $m_{\boldsymbol{\mathcal{B}}} = r_1 r_2 r_3, m_{\mathbf{D}_k} = (p_k - r_k) r_k$, and $k = 1, 2, 3$.
Solve $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^m} \|y - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|_2^2$.
Partition $\widehat{\boldsymbol{\gamma}}$ and assign each part to $\widehat{\boldsymbol{\mathcal{B}}}, \widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2, \widehat{\mathbf{D}}_3$, respectively,
    $$\mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) := \widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} = \widehat{\boldsymbol{\gamma}}_{[1:m_{\boldsymbol{\mathcal{B}}}]},$$
    $$\mathrm{vec}(\widehat{\mathbf{D}}_k) := \widehat{\boldsymbol{\gamma}}_{\mathbf{D}_k} = \widehat{\boldsymbol{\gamma}}_{\left[\left(m_{\boldsymbol{\mathcal{B}}}+\sum_{k'=1}^{k-1} m_{\mathbf{D}_{k'}}+1\right):\left(m_{\boldsymbol{\mathcal{B}}}+\sum_{k'=1}^{k} m_{\mathbf{D}_{k'}}\right)\right]}, \quad k = 1, 2, 3.$$
Let $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$. Evaluate
    $$\widehat{\boldsymbol{\mathcal{A}}} = [\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{L}}_1, \widehat{\mathbf{L}}_2, \widehat{\mathbf{L}}_3]\!], \quad \widehat{\mathbf{L}}_k = \left(\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widehat{\mathbf{D}}_k\right)\left(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right)^{-1}, \quad k = 1, 2, 3.$$
```

2.4 A Sketching Perspective of ISLET

While one of the main focuses of this article is on low-rank tensor regression, from a sketching perspective, ISLET can be seen as a special case of a more general algorithm that

Algorithm 2 Sparse Importance Sketching Low-rank Estimation for Tensors (Sparse ISLET): Order-3 Case

```
Input: sample $\{y_j, \boldsymbol{\mathcal{X}}_j\}_{j=1}^n$, Tucker rank $\boldsymbol{r} = (r_1, r_2, r_3)$, sparsity index $J_s \subseteq \{1, 2, 3\}$.
Evaluate $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{j=1}^n y_j \boldsymbol{\mathcal{X}}_j$.
Apply STAT-SVD on $\widetilde{\boldsymbol{\mathcal{A}}}$ with sparsity index $J_s$. Let the outcome be $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3$.
Let $\widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!]$ and evaluate the probing direction,
    $$\widetilde{\mathbf{V}}_k = \mathrm{QR}\left[\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right], \quad k = 1, 2, 3.$$
Construct
    $$\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n \times (r_1 r_2 r_3)}, \quad (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[i,:]} = \mathrm{vec}(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top),$$
    $$\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n \times (p_k r_k)}, \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[i,:]} = \mathrm{vec}\left(\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right)\widetilde{\mathbf{V}}_k\right).$$
Solve
    $$\widehat{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{r_1 r_2 r_3}, \quad \mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) = \operatorname*{arg\,min}_{\boldsymbol{\gamma} \in \mathbb{R}^{r_1 r_2 r_3}} \|y - \widetilde{\mathbf{X}}_{\mathbf{B}}\boldsymbol{\gamma}\|_2^2;$$
    $$\widehat{\mathbf{E}}_k \in \mathbb{R}^{p_k \times r_k}, \mathrm{vec}(\widehat{\mathbf{E}}_k) = \begin{cases} \arg\min_{\boldsymbol{\gamma}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2 + \lambda_k \sum_{j=1}^{p_k} \|\boldsymbol{\gamma}_{G_j^k}\|_2, & k \in J_s; \\ \arg\min_{\boldsymbol{\gamma}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2, & k \notin J_s. \end{cases}$$
Evaluate
    $$\widehat{\boldsymbol{\mathcal{A}}} = \left[\!\!\left[\widehat{\boldsymbol{\mathcal{B}}}; (\widehat{\mathbf{E}}_1(\widetilde{\mathbf{U}}_1^\top \widehat{\mathbf{E}}_1)^{-1}), (\widehat{\mathbf{E}}_2(\widetilde{\mathbf{U}}_2^\top \widehat{\mathbf{E}}_2)^{-1}), (\widehat{\mathbf{E}}_3(\widetilde{\mathbf{U}}_3^\top \widehat{\mathbf{E}}_3)^{-1})\right]\!\!\right].$$
```

broadly applies to high-dimensional statistical problems with dimension-reduced structure. In fact the three steps of the ISLET procedure are completely general and are summarized informally here:

Step 1 (Probing projection directions) For the tensor regression problem, we use the HOOI [36] or STAT-SVD [136] approach for finding the informative low-rank sub-spaces we project/sketch along. More generally if we let $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{j=1}^n y_j \boldsymbol{\mathcal{X}}_j$ where $\boldsymbol{\mathcal{X}}_j$ has ambient dimension p, we can define a general projection operator (with a slight abuse of notation) $\mathcal{P}_m(.) : \mathbb{R}^p \to \mathbb{R}^m$ indexed by low dimension m and let $\mathcal{S}(\widetilde{\boldsymbol{\mathcal{A}}})$ be the m-dimensional subspace of $\mathbb{R}^p$ determined by performing $\mathcal{P}_m(\widetilde{\boldsymbol{\mathcal{A}}})$.

Step 2 (Estimation in subspaces) The second step involves first projecting the data $\boldsymbol{\mathcal{X}}$ on to the subspace $\mathcal{S}(\widetilde{\boldsymbol{\mathcal{A}}})$, specifically $\widetilde{\mathbf{X}} = \mathcal{P}_{\mathcal{S}(\widetilde{\boldsymbol{\mathcal{A}}})}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{n \times m}$. Then perform regression or other procedures of choice using the sketched data $\widetilde{\mathbf{X}}$ to determine the dimension-reduced parameter $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^m$.

Step 3 (Embedding to high-dimensional space) Finally, we need to project the estimator back to the high-dimensional space $\mathbb{R}^p$ by applying an equivalent to the inverse of the projection operator $\mathcal{P}^{-1}_{\mathcal{S}(\widetilde{\boldsymbol{\mathcal{A}}})} : \mathbb{R}^m \to \mathbb{R}^p$. For low-rank tensor regression we require the formula (9).

The description above illustrates that the idea of ISLET is applicable to more general high-dimensional problems with dimension-reduced structure. In fact, the well-regarded *sure independence screening* in high-dimensional sparse linear regression [44, 129] can be seen as a special case of this idea. To be specific, consider the high-dimensional linear regression model,

$$y_i = X_{[i,:]}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\boldsymbol{\beta}$ is the m-sparse vector of interests and $y_i \in \mathbb{R}$ and $X^\top_{[i,:]} \in \mathbb{R}^p$ are observable response and covariate. Then the m-dimensional subspace $\mathcal{S}(\widetilde{\boldsymbol{\beta}})$ in Step 1 can be the co-ordinates corresponding to the m largest entries of $\widetilde{\boldsymbol{\beta}} = \sum_{i=1}^n X^\top_{[i,:]} y_i$; Step 2 corresponds to the dimension reduced least squares in sure independence screening; the inverse operator in Step 3 is simply filling in 0's in the co-ordinates that do not correspond to $\mathcal{S}(\widetilde{\boldsymbol{\beta}})$. In addition, this idea applies more broadly to problems such as matrix and tensor completion. One of the novel contributions of this article is finding suitable projection and inverse operators for low-rank tensors.

We can also contrast this approach with prior approaches that involve randomized sketching [38, 100, 102]. These prior approaches showed that the randomized sketching may lose data substantially, increases the variance, and yield sub-optimal result for many statistical problems. There are two key differences with how we exploit sketching in our context: (1) we sketch along the parameter directions of $\boldsymbol{\mathcal{X}}$, reducing the data from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times m}$; whereas approaches in [38, 100, 102] sketch along the sample directions, reducing the data from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{m\times p}$, which reduces the effective sample size from n to m; (2) secondly and most importantly rather than using the randomized sketching that is *unsupervised* without the response y, our importance sketching is *supervised* that is obtained using both the response y and covariates $\boldsymbol{\mathcal{X}}$. Then we sketch along the subspace $\mathcal{S}(\widetilde{\boldsymbol{\mathcal{A}}})$ which contains information on the low-dimensional structure of the parameter $\mathcal{A}$. This is why our general procedure has both desirable statistical and computational properties.

3 Oracle Inequalities

In this section, we provide general oracle inequalities without focusing on specific design, which provides a general guideline for the theoretical analyses of our ISLET procedure. We first introduce a quantification of the errors in sketching directions obtained in the first step of ISLET. Let $\mathbf{V}_k \in \mathbb{O}_{r_{k+1}r_{k+2}, r_k}$ be the right singular subspace of $\mathcal{M}_k(\boldsymbol{\mathcal{S}})$, where $\boldsymbol{\mathcal{S}}$ is the core tensor in the Tucker decomposition of $\boldsymbol{\mathcal{A}}$: $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!]$. By Lemma 1 in the supplementary material [137],

$$\begin{aligned} \mathbf{W}_1 := (\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{V}_1 \in \mathbb{O}_{p_2p_3,r_1}, \quad \mathbf{W}_2 := (\mathbf{U}_3 \otimes \mathbf{U}_1)\mathbf{V}_2 \in \mathbb{O}_{p_1p_3,r_2}, \\ \text{and} \quad \mathbf{W}_3 := (\mathbf{U}_2 \otimes \mathbf{U}_1)\mathbf{V}_3 \in \mathbb{O}_{p_1p_2,r_3} \end{aligned} \tag{18}$$

are the right singular subspaces of $\mathcal{M}_1(\boldsymbol{\mathcal{A}}), \mathcal{M}_2(\boldsymbol{\mathcal{A}})$, and $\mathcal{M}_3(\boldsymbol{\mathcal{A}})$, respectively. Recall that we initially estimate $\mathbf{U}_k$ and $\mathbf{V}_k$ by $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{V}}_k$, respectively in Step 1 of ISLET. Define

$$\widetilde{\mathbf{W}}_1 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1, \quad \widetilde{\mathbf{W}}_2 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_2, \quad \text{and} \quad \widetilde{\mathbf{W}}_3 = (\widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_3$$

in parallel to (18). Intuitively speaking, $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{W}}_k\}_{k=1}^3$ can be seen as the initial sample approximations for $\{\mathbf{U}_k, \mathbf{W}_k\}_{k=1}^3$. Therefore, we quantify the *sketching direction error* by

$$\theta := \max_{k=1,2,3} \left\{ \| \sin\Theta(\widetilde{\mathbf{U}}_k, \mathbf{U}_k)\|, \| \sin\Theta(\widetilde{\mathbf{W}}_k, \mathbf{W}_k)\| \right\}. \tag{19}$$

Next, we provide the oracle inequality via θ for ISLET under regular and sparse settings, respectively in the next two subsections.

3.1 Regular Tensor Regression and Oracle Inequality

In order to study the theoretical properties of the proposed procedure, we need to introduce another representation of the original model (1). Decompose the vectorized parameter $\boldsymbol{\mathcal{A}}$ as follows,

$$\begin{aligned} \mathrm{vec}(\boldsymbol{\mathcal{A}}) =& P_{\widetilde{\mathbf{U}}}\mathrm{vec}(\boldsymbol{\mathcal{A}}) + P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}}) \\ =& P_{\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1}\mathrm{vec}(\boldsymbol{\mathcal{A}}) + P_{\mathcal{R}_1(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp})}\mathrm{vec}(\boldsymbol{\mathcal{A}}) + P_{\mathcal{R}_2(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp})}\mathrm{vec}(\boldsymbol{\mathcal{A}}) \\ & + P_{\mathcal{R}_3(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp})}\mathrm{vec}(\boldsymbol{\mathcal{A}}) + P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}}) \\ =& (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) + \mathcal{R}_1(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp})\mathrm{vec}(\widetilde{\mathbf{D}}_1) + \mathcal{R}_2(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp})\mathrm{vec}(\widetilde{\mathbf{D}}_2) \\ & + \mathcal{R}_3(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp})\mathrm{vec}(\widetilde{\mathbf{D}}_3) + P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}}). \end{aligned} \tag{20}$$

(See the proof of Theorem 2 for a detailed derivation of (20)). Here,

$$\widetilde{\mathbf{U}} = \left[\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1 \;\; \mathcal{R}_1(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp}) \;\; \mathcal{R}_2\left(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp}\right) \;\; \mathcal{R}_3\left(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp}\right) \right],$$

$$\widetilde{\boldsymbol{\mathcal{B}}} := \left[\!\left[\boldsymbol{\mathcal{A}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top \right]\!\right] \in \mathbb{R}^{r_1r_2r_3} \quad \text{and} \quad \widetilde{\mathbf{D}}_k := \widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}})\widetilde{\mathbf{W}}_k \in \mathbb{R}^{(p_k - r_k)\times r_k}$$

are the singular subspace of the "Cross structure" and the low-dimensional projections of $\boldsymbol{\mathcal{A}}$ onto the "body" and "arms" formed by sketching directions $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k\}_{k=1}^3$, respectively (See Figure 4 for an illustration of $\widetilde{\mathbf{U}}$, $\widetilde{\boldsymbol{\mathcal{B}}}$, and $\widetilde{\mathbf{V}}_k$). Due to different alignments, the i-th row of $\{\mathbf{W}_k \otimes \mathbf{U}_{k\perp}\}_{k=1}^3$ does not necessarily correspond to the i-th entry of $\mathrm{vec}(\boldsymbol{\mathcal{A}})$ for

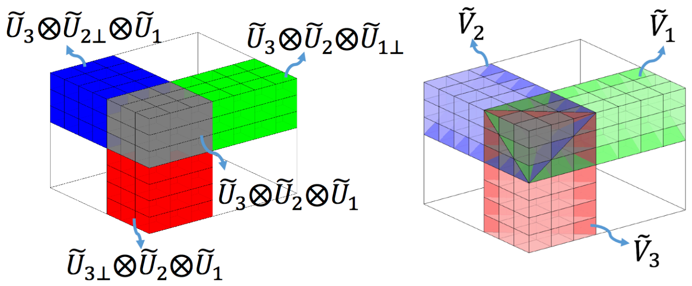

$\widetilde{U}_3 \otimes \widetilde{U}_{2\perp} \otimes \widetilde{U}_1$
$\widetilde{U}_3 \otimes \widetilde{U}_2 \otimes \widetilde{U}_{1\perp}$
$\widetilde{U}_3 \otimes \widetilde{U}_2 \otimes \widetilde{U}_1$
$\widetilde{U}_{3\perp} \otimes \widetilde{U}_2 \otimes \widetilde{U}_1$
$\widetilde{V}_2$
$\widetilde{V}_1$
$\widetilde{V}_3$

Figure 4: Illustration of Decomposition (20). Here, we assume $\widetilde{\mathbf{U}}_k^\top = [\mathbf{I}_{r_k} \ \mathbf{0}_{r_k \times (p_k - r_k)}]$, $k = 1, 2, 3$, for a better visualization. The gray, green, blue, and red cubes represent the subspaces of $\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1$, $\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_{1\perp}$, $\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_{2\perp} \otimes \widetilde{\mathbf{U}}_1$, $\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1$. The gray cube also corresponds to the projected parameters $\widetilde{\boldsymbol{\mathcal{B}}}$; matricizations of green, blue and red cubes correspond to the projected parameters $\widetilde{\mathbf{U}}_{1\perp}^\top \mathcal{M}_1(\boldsymbol{\mathcal{A}})(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)$, $\widetilde{\mathbf{U}}_{2\perp}^\top \mathcal{M}_2(\boldsymbol{\mathcal{A}})(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_1)$, and $\widetilde{\mathbf{U}}_{3\perp}^\top \mathcal{M}_3(\boldsymbol{\mathcal{A}})(\widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)$, respectively. The three plains in the right panel correspond to the subspace of $\widetilde{\mathbf{V}}_1$, $\widetilde{\mathbf{V}}_2$, and $\widetilde{\mathbf{V}}_3$, respectively.

all $1 \leq i \leq p_1 p_2 p_3$. We thus permute the rows of $\{\widetilde{\mathbf{W}}_k \otimes \widetilde{\mathbf{U}}_{k\perp}\}_{k=1}^3$ to match each row of $\mathcal{R}_k(\widetilde{\mathbf{W}}_k \otimes \widetilde{\mathbf{U}}_{k\perp})$ to the corresponding entry in $\text{vec}(\boldsymbol{\mathcal{A}})$. The formal definition of the row-wise permutation operator $\mathcal{R}_k$ is rather clunky and postponed to Section A in the supplementary materials. Intuitively speaking, $P_{\widetilde{\mathbf{U}}}\text{vec}(\boldsymbol{\mathcal{A}})$ represents the projection of $\boldsymbol{\mathcal{A}}$ onto to the Cross structure and $P_{\widetilde{\mathbf{U}}_\perp}\text{vec}(\boldsymbol{\mathcal{A}})$ can be seen as a residual. If the estimates $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{W}}_k\}_{k=1}^3$ are close enough to $\{\mathbf{U}_k, \mathbf{W}_k\}_{k=1}^3$, i.e., θ defined in (19) is small, we expect that the residual $P_{\widetilde{\mathbf{U}}_\perp}\text{vec}(\boldsymbol{\mathcal{A}})$ has small amplitude.

Based on (20), we can re-write the original regression model (1) into the following partial regression model,

$$\begin{aligned} y_j =& (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[j,:]}\text{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) + \sum_{k=1}^{3} (\widetilde{\mathbf{X}}_{\mathbf{D}_k})_{[j,:]}\text{vec}(\widetilde{\mathbf{D}}_k) + \text{vec}(\boldsymbol{\mathcal{X}}_j)^\top P_{\widetilde{\mathbf{U}}_\perp}\text{vec}(\boldsymbol{\mathcal{A}}) + \varepsilon_j \\ =& \widetilde{\mathbf{X}}_{[j,:]}\widetilde{\gamma} + \widetilde{\varepsilon}_j, \quad j = 1, \ldots, n. \end{aligned} \tag{21}$$

(See the proof of Theorem 2 for a detailed derivation of (21).) Here,

- $\widetilde{\varepsilon}_j = \text{vec}(\boldsymbol{\mathcal{X}}_j)^\top P_{\widetilde{\mathbf{U}}_\perp}\text{vec}(\boldsymbol{\mathcal{A}}) + \varepsilon_j$ is the oracle noise; $\widetilde{\boldsymbol{\varepsilon}} = (\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n)^\top$;
- $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}, \widetilde{\mathbf{X}}_{\mathbf{D}_k}$ are sketching covariates introduced in Equation (6);
- $\widetilde{\gamma} = \left[\text{vec}(\widetilde{\boldsymbol{\mathcal{B}}})^\top, \text{vec}(\widetilde{\mathbf{D}}_1)^\top, \text{vec}(\widetilde{\mathbf{D}}_2)^\top, \text{vec}(\widetilde{\mathbf{D}}_3)^\top\right]^\top = \widetilde{\mathbf{U}}^\top\text{vec}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^m$ is the dimension-reduced parameter.

(21) reveals the essence of the least squares estimator (7) in the ISLET procedure – the outcomes of (7) and (8), i.e., $\widehat{\boldsymbol{\mathcal{B}}}$ and $\widehat{\mathbf{D}}_k$, are sample-based estimates of $\widetilde{\boldsymbol{\mathcal{B}}}$ and $\widetilde{\mathbf{D}}_k$. Finally,

based on the detailed algebraic calculation in Step 3 and the proof of Theorem 2,

$$\boldsymbol{\mathcal{A}} = \left\llbracket \widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2, \widetilde{\mathbf{L}}_3 \right\rrbracket, \quad \widetilde{\mathbf{L}}_k = \left(\widetilde{\mathbf{U}}_k \widetilde{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widetilde{\mathbf{D}}_k\right) \left(\widetilde{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right)^{-1}. \tag{22}$$

(22) is essentially a higher-order version of the Schur complement formula (also see [20]). Finally, we apply the plug-in estimator to obtain the final estimator $\widehat{\boldsymbol{\mathcal{A}}}$ (Equation (9) in Step 3 of the ISLET procedure).

Based on previous discussions, it can be seen that the estimation error of the original tensor regression is driven by the error of the least squares estimator $\widehat{\gamma}$, i.e., $\|(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\boldsymbol{\varepsilon}}\|_2^2$. We have the following oracle inequality for the proposed ISLET procedure.

Theorem 2 (Oracle Inequality of Regular Tensor Estimation: Order-3 Case)**.** *Suppose $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ has Tucker rank-(r_1, r_2, r_3) tensor and $\widehat{\boldsymbol{\mathcal{A}}}$ is the outcome of Algorithm 1. Assume the sketching directions $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k\}_{k=1}^3$ satisfy $\theta < 1/2$ (see* (19) *for the definition of θ) and $\left\|\widehat{\mathbf{D}}_k (\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k)^{-1}\right\| \leq \rho$. We don't impose other specific assumptions on $\boldsymbol{\mathcal{X}}_i$ and ε_i. Then, we have*

$$\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \leq (1 + C(\theta + \rho)) \left\|(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\boldsymbol{\varepsilon}}\right\|_2^2$$

for uniform constant $C > 0$ that does not rely on any other parameters.

Proof. See Appendix F.1 for a complete proof. In particular, the proof contains three major steps. After introducing a number of notation, we first transform the original regression model to the partial regression model (21), then rewrite the upper bound $\|(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\boldsymbol{\varepsilon}}\|_2^2$ to $\|\widehat{\boldsymbol{\mathcal{B}}} - \widetilde{\boldsymbol{\mathcal{B}}}\|_{\mathrm{HS}}^2 + \sum_{k=1}^3 \|\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k\|_F^2$. Next, we introduce a factorization of $\boldsymbol{\mathcal{A}}$ in parallel with the one of $\widehat{\boldsymbol{\mathcal{A}}}$, based on which the loss $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}$ is decomposed into eight terms. Finally, we introduce a novel deterministic error bound for the "Cross scheme" (Lemma 3 in the supplementary material [137]; also see [135]), carefully analyze each term in the decomposition of $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}$, and finalize the proof. □

Theorem 2 shows that once the sketching directions $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are reasonably accurate, the estimation error for $\widehat{\boldsymbol{\mathcal{A}}}$ will be close to the error of partial linear regression in Equation (21). This bound is general and deterministic, which can be used as a key step in more specific settings of low-rank tensor regression.

3.2 Sparse Tensor Regression and Oracle Inequality

Next, we study the oracle performance of the proposed procedure for sparse tensor regression, where $\boldsymbol{\mathcal{A}}$ further satisfies the sparsity constraint (3). As in the previous section, we decompose the vectorized parameter as

$$\begin{aligned}\mathrm{vec}(\boldsymbol{\mathcal{A}}) =& P_{\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1} \mathrm{vec}(\boldsymbol{\mathcal{A}}) + P_{(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)_\perp} \mathrm{vec}(\boldsymbol{\mathcal{A}}) \\ =& (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_3) \mathrm{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) + P_{(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)_\perp} \mathrm{vec}(\boldsymbol{\mathcal{A}});\end{aligned} \tag{23}$$

$$\begin{aligned}\operatorname{vec}(\boldsymbol{\mathcal{A}}) =& P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes \mathbf{I}_{p_k})}\operatorname{vec}(\boldsymbol{\mathcal{A}}) + P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes \mathbf{I}_{p_k})_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}}) \\ =& \mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes \mathbf{I}_{p_k})\operatorname{vec}(\widetilde{\mathbf{E}}_k) + P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes \mathbf{I}_{p_k})_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}}), \quad k=1,2,3.\end{aligned} \tag{24}$$

Here,

$$\begin{aligned}&\widetilde{\boldsymbol{\mathcal{B}}} := [\![\boldsymbol{\mathcal{A}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!] \in \mathbb{R}^{r_1r_2r_3}; \\ &\widetilde{\mathbf{E}}_k := \mathcal{M}_k\left(\boldsymbol{\mathcal{A}} \times_{(k+1)} \widetilde{\mathbf{U}}_{k+1}^\top \times_{(k+2)} \widetilde{\mathbf{U}}_{k+2}^\top\right)\widetilde{\mathbf{V}}_k \in \mathbb{R}^{p_k\times r_k}, \quad k=1,2,3,\end{aligned} \tag{25}$$

are the low-dimensional projections of $\boldsymbol{\mathcal{A}}$ onto the importance sketching directions. Since $\{\mathbf{U}_k, \mathbf{W}_k\}$ are the left and right singular subspaces of $\mathcal{M}_k(\boldsymbol{\mathcal{A}})$, we can show $P_{(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})$ and $P_{\mathcal{R}_k(\mathbf{W}_k\otimes\mathbf{I}_{p_k})_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})$ are zeros. Thus if the estimates $\{\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{W}}_k\}_{k=1}^3$ are sufficiently accurate, i.e., θ defined in Eq. (19) is small, we expect that the residuals $P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})$ and $P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})$ have small amplitudes. Then, based on a more detailed calculation in the proof of Theorem 3, the model of sparse and low-rank tensor regression $y_j = \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}}\rangle + \varepsilon_j$ can be rewritten as the following partial linear regression,

$$y_j = (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[j,:]}\operatorname{vec}(\widetilde{\mathbf{B}}) + (\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}})_j, \tag{26}$$

$$y_j = (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[j,:]}\operatorname{vec}(\widetilde{\mathbf{E}}_k) + (\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k})_j, \quad k=1,2,3. \tag{27}$$

Here, $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}$ and $\widetilde{\mathbf{X}}_{\mathbf{E}_k}$ are the covariates defined in Equation (13) and $\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}} = ((\widetilde{\varepsilon}_{\boldsymbol{\mathcal{B}}})_1, \ldots, (\widetilde{\varepsilon}_{\boldsymbol{\mathcal{B}}})_n)^\top$, $\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k} = ((\widetilde{\varepsilon}_{\mathbf{E}_k})_1, \ldots, (\widetilde{\varepsilon}_{\mathbf{E}_k})_n)^\top$ are oracle noises defined as

$$\begin{aligned}(\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}})_j &= \left\langle \operatorname{vec}(\boldsymbol{\mathcal{X}}_j), P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})\right\rangle + \varepsilon_j \\ \text{and}\quad (\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k})_j &= \left\langle \operatorname{vec}(\boldsymbol{\mathcal{X}}_j), P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}\operatorname{vec}(\boldsymbol{\mathcal{A}})\right\rangle + \varepsilon_j.\end{aligned} \tag{28}$$

Therefore, the Step 2 of sparse ISLET can be interpreted as the estimation of $\widetilde{\boldsymbol{\mathcal{B}}}$ and $\widetilde{\mathbf{E}}_k$.

We apply regular least squares to estimate $\widetilde{\boldsymbol{\mathcal{B}}}$ and $\widetilde{\mathbf{E}}_k$ for $k \notin J_s$. For any sparse mode $k \in J_s$, $\widetilde{\mathbf{E}}_k$ are group sparse due to the definition (25) and the assumption that $\mathbf{U}_k$ are row-wise sparse. Specifically, $\widetilde{\mathbf{E}}_k$ satisfies

$$\left\|\operatorname{vec}(\widetilde{\mathbf{E}}_k)\right\|_{0,2} := \sum_{i=1}^{p_k} 1_{\left\{\left(\operatorname{vec}(\widetilde{\mathbf{E}}_k)\right)_{G_i^k}\neq 0\right\}} \leq s_k, \tag{29}$$

where

$$G_i^k = \{i, i+p_k, \ldots, i+p_k(r_k-1)\}, \quad i=1,\ldots,p_k, \quad \forall k \in J_s,$$

is a partition of $\{1,\ldots,p_kr_k\}$ (see the proof for Theorem 3 for a more detailed argument for (29)). By detailed calculations in Step 3 of the proof for Theorem 2, one can verify that

$$\boldsymbol{\mathcal{A}} = \left[\!\!\left[\widetilde{\boldsymbol{\mathcal{B}}}, (\widetilde{\mathbf{E}}_1(\widetilde{\mathbf{U}}_1^\top\widetilde{\mathbf{E}}_1)^{-1}), (\widetilde{\mathbf{E}}_2(\widetilde{\mathbf{U}}_2^\top\widetilde{\mathbf{E}}_2)^{-1}), (\widetilde{\mathbf{E}}_3(\widetilde{\mathbf{U}}_3^\top\widetilde{\mathbf{E}}_3)^{-1})\right]\!\!\right].$$

Then the finally sparse ISLET estimator $\widehat{\boldsymbol{\mathcal{A}}}$ in (17) can be seen as the plug-in estimator.

To ensure that the group Lasso estimator in (15) provides a stable estimation for the proposed procedure, we introduce the following group restricted isometry condition, which can also be seen as an extension of restricted isometry property (RIP), a commonly used condition in compressed sensing and high-dimensional linear regression literature [26].

Condition 1. *We say a matrix* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *satisfies the group restricted isometry property (GRIP) with respect to partition* $G_1, \ldots, G_m \subseteq \{1, \ldots, p\}$*, if there exists* $\delta > 0$ *such that*

$$n(1-\delta)\|\mathbf{v}\|_2^2 \le \|\mathbf{X}\mathbf{v}\|_2^2 \le n(1+\delta)\|\mathbf{v}\|_2^2 \tag{30}$$

for all group-wise sparse vector v *satisfying* $\sum_{k=1}^m 1_{\{\mathbf{v}_{G_k}\neq 0\}} \le s$.

We still use θ defined in Eq. (19) to characterize the sketching direction errors. The following oracle inequality holds for sparse tensor regression with importance sketching.

Theorem 3 (Oracle Inequality for Sparse Tensor Regression: Order-3 Case)**.** *Consider the sparse low-rank tensor regression* (1) (3)*. Suppose* $\theta < 1/2$*, the importance sketching covariates* $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}$ *and* $\widetilde{\mathbf{X}}_{\mathbf{E}_k}$ $(k \notin J_s)$ *are non-singular. For any* $k \in J_s$, $\widetilde{\mathbf{X}}_{\mathbf{E}_k}$ *satisfies group restricted isometry property (Condition 1) with respect to partition* $G_1^k, \ldots, G_{p_k}^k$ *in* (16) *and* $\delta < 1/3$*. We apply the proposed Algorithm 2 with group Lasso penalty*

$$\eta_k = C_1 \max_{i=1,\ldots,p_k} \left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k,[:,G_i^k]})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2$$

for $k \in J_s$ *and some constant* $C_1 \ge 3$*. We also assume* $\|\widetilde{\mathbf{U}}_{k\perp}^\top \widehat{\mathbf{E}}_k(\widetilde{\mathbf{U}}_k^\top \widehat{\mathbf{E}}_k)^{-1}\| \le \rho$*. Then,*

$$\begin{aligned}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 &\le (1 + C_2 s(\theta+\rho))\Bigg(\left\|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{B}}\right\|_2^2 \\ &+ \sum_{k\notin J_s}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1}\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2 + C_3\sum_{k\in J_s} s_k \cdot \max_{i=1,\ldots,p_k}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k,[:,G_i^k]})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n\right\|_2^2\Bigg).\end{aligned} \tag{31}$$

Proof. See Appendix F.2. □

Remark 3. *In the oracle error bound* (31), $\|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{B}}\|_2^2$, $\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1}\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{B}}\right\|_2^2$*, and* $s_k \max_{i=1,\ldots,p_k}\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k,[:,G_i^k]})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n\|_2^2$ *correspond to the estimation errors of* $\widehat{\boldsymbol{\mathcal{B}}}$, $\widehat{\mathbf{E}}_k$ *of the non-sparse mode, and* $\widehat{\mathbf{E}}_k$ *of sparse mode, respectively. When the group restricted isometry property (Condition 1) is replaced by group restricted eigenvalue condition (see, e.g., [79]), a similar result to Theorem 3 can be derived.*

4 Fast Low-rank Tensor Regression via ISLET

We further study the low-rank tensor regression with Gaussian ensemble design, i.e., $\boldsymbol{\mathcal{X}}_i$ has i.i.d. standard normal entries. This has been considered a benchmark setting for low-rank tensor/matrix recovery literature [25, 29]. For convenience, we denote $\boldsymbol{p} = (p_1, p_2, p_3), \boldsymbol{r} = (r_1, r_2, r_3)$, $p = \max\{p_1, p_2, p_3\}$, and $r = \max\{r_1, r_2, r_3\}$. We discuss the regular low-rank and sparse low-rank tensor regression in the next two subsections, respectively.

4.1 Regular Low-rank Tensor Regression with ISLET

We have the following theoretical guarantee for ISLET under Gaussian ensemble design.

Theorem 4 (Upper bound for tensor regression via ISLET)**.** *Consider the tensor regression model* (1)*, where $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times p_2\times p_3}$ is Tucker rank-(r_1, r_2, r_3), $\boldsymbol{\mathcal{X}}_i$ has i.i.d. standard normal entries, and $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$. Denote $\widetilde{\sigma}^2 = \|\boldsymbol{\mathcal{A}}\|_{\rm HS}^2 + \sigma^2$, $\lambda_0 = \min_k \lambda_k, \lambda_k = \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))$, $\kappa = \max_k \|\mathcal{M}_k(\boldsymbol{\mathcal{A}})\|/\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))$, and $m = r_1r_2r_3 + \sum_{k=1}^3(p_k - r_k)r_k$. If $n_1 \wedge n_2 \geq \frac{C\widetilde{\sigma}^2(p^{3/2}+\kappa pr)}{\lambda_0^2}$, then the sample-splitting ISLET estimator (see the forthcoming Remark 5) satisfies*

$$\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\rm HS}^2 \leq \frac{m}{n_2}\left(\sigma^2 + \frac{C_1\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}\right)\left(1 + C_2\sqrt{\frac{\log p}{m}} + C_3\sqrt{\frac{m\widetilde{\sigma}^2}{(n_1\wedge n_2)\lambda_0^2}}\right)$$

with probability at least $1 - p^{-C_4}$.

Proof. See Section F.3 for details. Specifically, we first derive the estimation error upper bounds for sketching directions $\widetilde{\mathbf{U}}_k$ via the deterministic error bound of HOOI [138]. Then we apply concentration inequalities to obtain upper bounds for $\left\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{\varepsilon}}\right\|_2^2$ and $\|\widehat{\mathbf{D}}_k(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\|$ for $k = 1, 2, 3$. Finally, the oracle inequality of Theorem 2 leads to the desired upper bound. □

Remark 4 (Sample Complexity)**.** *In Theorem 4, we show that as long as the sample size $n = \Omega(p^{3/2}r + pr^2)$, ISLET achieves consistent estimation under regularity conditions. This sample complexity outperforms many computationally feasible algorithms in previous literature, e.g., $n = \Omega(p^2 r polylog(p))$ in projected gradient descent [29], sum of nuclear norm minimization [117], and square norm minimization [91]. To the best of our knowledge, ISLET is the first computationally efficient algorithm that achieves this sample complexity result.*

On the other hand, [91] showed that the direct nonconvex Tucker rank minimization, a computationally infeasible method, can do exact recovery with $O(pr + r^3)$ linear measurements in the noiseless setting. [13] showed that if tensor parameter $\boldsymbol{\mathcal{A}}$ is CP rank-r, the linear system $y_j = \langle\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}_j\rangle, j = 1, \ldots, n$ has a unique solution with probability one if one *has $O(pr)$ measurements. It remains an open question whether the sample complexity of $n = \Omega(p^{3/2}r + pr^2)$ is necessary for all computationally efficient procedures.*

Remark 5 (Sample splitting)**.** *The direct analysis for the proposed ISLET in Algorithm 1 is technically involved, among which one major difficulty is the dependency between the sketching directions $\widetilde{\mathbf{U}}_k$ obtained in Step 1 and the regression noise $\widetilde{\boldsymbol{\varepsilon}}$ in Step 2. To overcome this difficulty, we choose to analyze a modified procedure with the sample splitting scheme: we randomly split all n samples into two sets with cardinality n_1 and n_2, respectively. Then we use the first set of n_1 samples to construct the covariance tensor $\widetilde{\boldsymbol{\mathcal{A}}}$ (Step 1) and use the second set of n_2 samples to evaluate the importance sketching covariates (Step 2). As illustrated by numerical studies in Section 5, such a scheme is mainly for technical purposes and is not necessary in practice. Simulations suggest that it is preferable to use all samples $\{y_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n$ for both constructing the initial estimate $\widetilde{\boldsymbol{\mathcal{A}}}$ and performing linear regression on sketching covariates.*

We further consider the statistical limits for low-rank tensor regression with Gaussian ensemble. Consider the following class of general low-rank tensors,

$$\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}} = \left\{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times p_2\times p_3} : \text{Tucker rank}(\boldsymbol{\mathcal{A}}) \leq (r_1, r_2, r_3)\right\}. \tag{32}$$

The following minimax lower bound holds for all low-rank tensors in $\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}$.

Theorem 5 (Minimax Lower Bound)**.** *If $n > m+1$, the following non-asymptotic lower bound in estimation error hold,*

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \frac{m}{n-m-1}\cdot\sigma^2. \tag{33}$$

If $n \leq m+1$,

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 = +\infty. \tag{34}$$

Proof. See Appendix F.4. □

Combining Theorems 4 and 5, we can see that as long as the sample size satisfies $\frac{m\widetilde{\sigma}^2}{n_1\lambda_0^2} = o(1)$, $\frac{m(p_1+p_2+p_3)\widetilde{\sigma}^4}{n_1n_2\lambda_0^2} = o(\sigma^2)$, and $n_2 = (1+o(1))n$, the statistical loss of the proposed method is sharp with matching constant to the lower bound.

Remark 6 (Matrix ISLET vs. Previous Matrix Recovery Methods)**.** *If the order of tensor reduces to two, the tensor regression becomes the well-regarded* low-rank matrix recovery *in literature [25, 104]:*

$$y_i = \langle \mathbf{X}_i, \mathbf{A}\rangle + \varepsilon_i, \quad i = 1, \ldots, n.$$

Here, $\mathbf{A} \in \mathbb{R}^{p_1\times p_2}$ is the unknown rank-r target matrix, $\{\mathbf{X}_i\}_{i=1}^n$ are design matrices, and $\varepsilon_i \sim N(0, \sigma^2)$ are noises. The low-rank matrix recovery, including its instances such as *phase retrieval [23], has been widely considered in recent literature. Various methods, such as nuclear norm minimization [24, 104], projected gradient descent [115], singular value thresholding [15], Procrustes flow [119], etc, have been introduced and both the theoretical and computational performances have been extensively studied. By similar proof of Theorem 4, the following upper bound for matrix ISLET estimator $\widehat{\mathbf{A}}$ (Algorithm 4 in the supplementary material [137])*

$$\left\|\widehat{\mathbf{A}}-\mathbf{A}\right\|_F^2 \leq \frac{m}{n_2}\left(\sigma^2+\frac{C_1\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}\right)\left(1+C_2\sqrt{\frac{\log p}{m}}+C_3\sqrt{\frac{m\widetilde{\sigma}^2}{(n_1\wedge n_2)\lambda_0^2}}\right)$$

can be established with high probability. Here, $m=(p_1+p_2-r)r$, $\lambda_0=\sigma_r(\mathbf{A}), \widetilde{\sigma}^2=\|\mathbf{A}\|_F^2+\sigma^2$. The lower bound similarly to Theorem 5 also holds.

4.2 Sparse Tensor Regression with Importance Sketching

We further consider the simultaneously sparse and low-rank tensor regression with Gaussian ensemble design. We have the following theoretical guarantee for sparse ISLET. Due to the same reason as for regular ISLET (see Remark 5), the sample splitting scheme is introduced in our technical analysis.

Theorem 6 (Upper Bounds for Sparse Tensor Regression via ISLET)**.** *Consider the tensor regression model* (1)*, where $\boldsymbol{\mathcal{A}}$ is simultaneously low-rank and sparse* (3)*, $\boldsymbol{\mathcal{X}}_i$ has i.i.d. standard Gaussian entries, and $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. Denote $\lambda_0=\min_k \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))$, $s_k=p_k$ if $k\notin J_s$, $m_s=r_1r_2r_3+\sum_{k\in J_s}s_k(r_k+\log p_k)+\sum_{k\notin J_s}p_kr_k$, and $\kappa=\max_k\|\mathcal{M}_k(\boldsymbol{\mathcal{A}})\|/\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))$. We apply the proposed Algorithm 2 with sample splitting scheme (see Remark 5) and group Lasso penalty $\eta_k=C_0\widetilde{\sigma}\sqrt{n_2(r_k+\log(p_k))}$. If* $\log(p_1)\asymp\log(p_2)\asymp\log(p_3)\asymp\log(p)$,

$$n_1\geq\frac{C_1\kappa^2\widetilde{\sigma}^2}{\lambda_0^2}\left(s_1s_2s_3\log(p)+\sum_{k=1}^3(s_k^2r_k^2+r_{k+1}^2r_{k+2}^2)\right),\quad n_2\geq\frac{C_2m_s\kappa^2\widetilde{\sigma}^2}{\lambda_0^2},$$

the output $\widehat{\boldsymbol{\mathcal{A}}}$ of sparse ISLET satisfies

$$\left\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2\leq\frac{C_3m_s}{n_2}\left(\sigma^2+\frac{C_4m_s\kappa^2\widetilde{\sigma}^2}{n_1}\right) \tag{35}$$

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

Proof. See Appendix F.5. □

We further consider the following class of simultaneously sparse and low-rank tensors,

$$\mathcal{A}_{\boldsymbol{p},\boldsymbol{r},\boldsymbol{s}}=\left\{\boldsymbol{\mathcal{A}}=[\![\boldsymbol{\mathcal{S}};\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3]\!]:\mathbf{U}_k\in\mathbb{O}_{p_k,r_k},\|\mathbf{U}_k\|_{0,2}\leq s_k,k\in J_s\right\}. \tag{36}$$

The following minimax lower bound of the estimation risk holds in this class.

Theorem 7 (Lower Bounds)**.** *There exists constant $C > 0$ such that whenever $m_s \geq C$, the following lower bound holds for any arbitrary estimator $\widehat{\boldsymbol{\mathcal{A}}}$ based on $\{\boldsymbol{\mathcal{X}}_i, y_i\}_{i=1}^n$,*

$$\inf_{\boldsymbol{\mathcal{A}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r},\boldsymbol{s}}} \mathbb{E} \left\| \widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{HS}}^2 \geq \frac{c m_s}{n} \sigma^2. \tag{37}$$

Proof. See Appendix F.6. □

Combining Theorems 6 and 7, we can see the proposed procedure achieves optimal rate of convergence if $\frac{m_s \|\boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}^2}{n_1 \sigma^2} = O(1)$ and $n_2 \asymp n$.

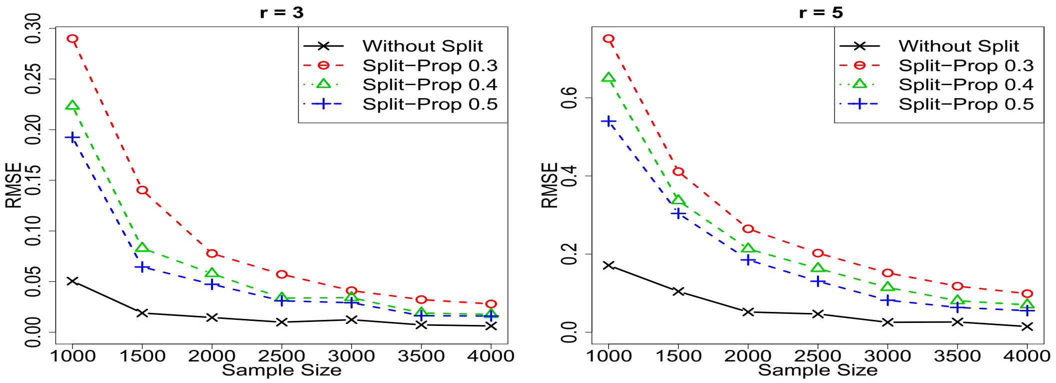
5 Numerical Analysis

In this section, we conduct a simulation study to investigate the numerical performance of ISLET. In each study, we construct sensing tensors $\boldsymbol{\mathcal{X}}_j \in \mathbb{R}^{p \times p \times p}$ with independent standard normal entries. In the non-sparse settings, using the Tucker decomposition we generate the core tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r \times r \times r}$ and $\mathbf{E}_k \in \mathbb{R}_{p,r}$ with i.i.d. Gaussian entries, the coefficient tensor $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{E}_1; \mathbf{E}_2; \mathbf{E}_3]\!]$; in the sparse settings, we construct $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{A}}$ in the same way and generate $\mathbf{E}_k$ as

$$(\mathbf{E}_k)_{[i,:]} = \begin{cases} (\bar{\mathbf{E}}_k)_{[j,:]}, & i \in \Omega_k, \text{ and } i \text{ is the } j\text{-th element of } \Omega_k; \\ 0, & i \notin \Omega_k, \end{cases}$$

where Ω_k is a uniform random subset of $\{1, \ldots, p\}$ with cardinality s_k and $\bar{\mathbf{E}}_k$ has s_k-by-r i.i.d. Gaussian entries. Finally, let the response $y_j = \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}} \rangle + \varepsilon_j, j = 1, 2, \ldots, n$, where $\varepsilon_j \overset{iid}{\sim} N(0, \sigma^2)$. We report both the average root mean squared error (RMSE) $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}||_{\mathrm{HS}}/||\boldsymbol{\mathcal{A}}||_{\mathrm{HS}}$ and the run time for each setting. Unless otherwise noted, the reported results are based on the average of 100 repeats and on a computer with Intel Xeon E5-2680 2.50GHz CPU. Additional simulation results of tuning-free ISLET and approximate low-rank tensor regression are collected in Sections D and E in the supplementary material [137].

Since we proposed to evaluate sketching directions and dimension-reduced regression (Steps 1 and 2 of Algorithm 1) both using the complete sample, but introduced a sample splitting scheme (Remark 5) to prove Theorems 4 and 6, we investigate how the sample splitting scheme affects the numerical performance of ISLET in this simulation setting. Let n vary from 1000 to 4000, $p = 10$, $r = 3, 5$, $\sigma = 5$. In addition to the original ISLET without splitting, we also implement sample-splitting ISLET, where a random $n_1 \approx \{\frac{3}{10}n, \frac{4}{10}n, \frac{5}{10}n\}$ samples are allocated for importance direction estimation (Step 1 of ISLET) and $n - n_1$ are allocated for dimension-reduced regression (Step 2 of ISLET). The results plotted in Figure 5 clearly show that the no-sample-splitting scheme yields much smaller estimation error than all sample-splitting approaches. Although the sample splitting scheme brings advantages for our theoretical analyses for ISLET, it is not necessary in practice. Therefore, we will only perform ISLET without sample splitting for the rest of the simulation studies.



(a) $r = 3$

(b) $r = 5$

Figure 5: No-splitting vs. splitting ISLET: n varies from 1000 to 4000, $p = 10$, $r = 3, 5$, $\sigma = 5$.

We also compare the performance of non-sparse ISLET with a number of contemporary methods, including non-convex projected gradient descent (non-convex PGD) [29], Tucker low-rank regression via alternating gradient descent (Tucker regression)[1] [77, 143], and convex regularization low-rank tensor recovery (convex regularization)[2] [78, 103, 117]. We implement all four methods for $p = 10$, but only the ISLET and non-convex projected PGD for $p = 50$, as the time cost of Tucker regression and convex regularization are beyond our computational limit if $p = 50$. Results for $p = 10$ and $p = 50$ are respectively plotted in Panels (a)(b) and Panels (c)(d) of Fig. 6. Plots in Fig. 6 (a) and (c) show that the RMSEs of ISLET, tucker tensor regression and non-convex PGD are close, and all of them are slightly better than the convex regularization method; Figure 6 (b) and (d) further indicate that ISLET is much faster than other methods – the advantage significantly increases as n and p grow. In particular, ISLET is about 10 times faster than non-convex PGD when $p = 50, n = 12000$. In summary, the proposed ISLET achieves similar statistical performance within in a significantly shorter time period comparing to the other state-or-

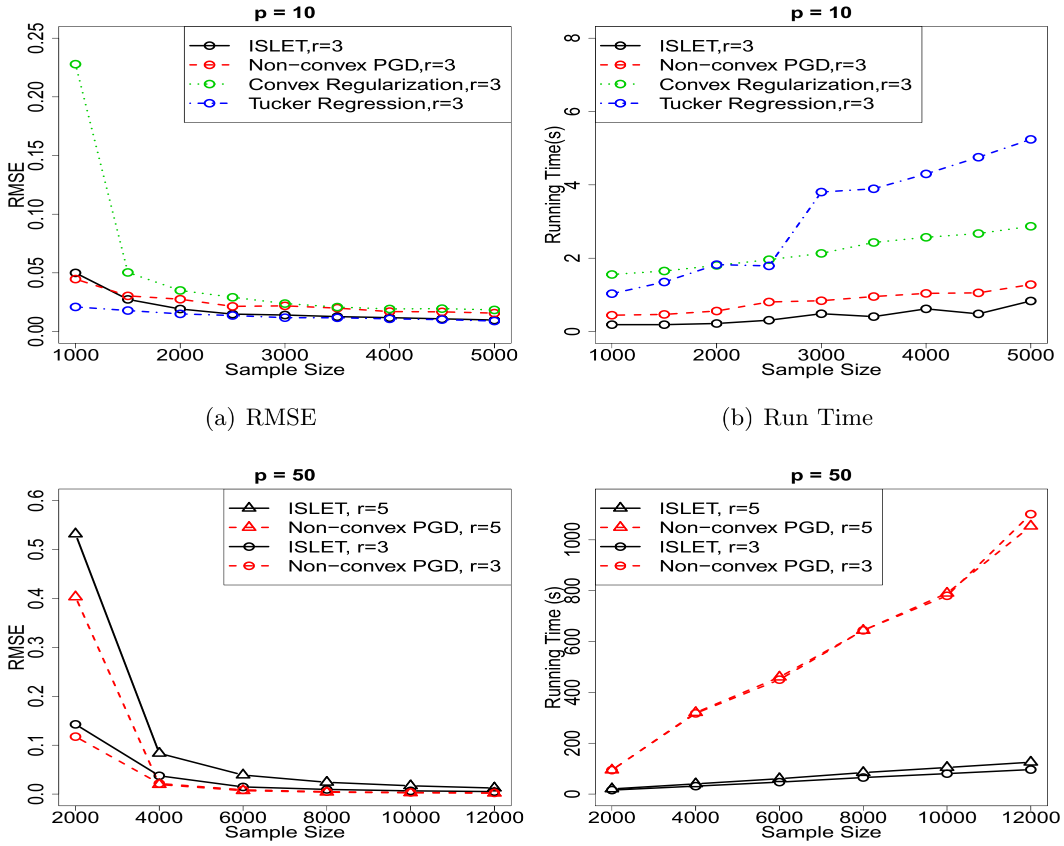
[1] Software package downloaded at `https://hua-zhou.github.io/TensorReg/`

[2] The convex regularization aims to minimize the following objective function

$$\sum_{i}^{n} \frac{1}{2n}(y_i - \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}} \rangle)^2 + \lambda \sum_{k=1}^{3} ||\mathcal{M}_k(\boldsymbol{\mathcal{A}})||_*.$$

Here, $\|\cdot\|_*$ is the matrix nuclear norm.

the-art methods.



(a) RMSE

(b) Run Time

(c) RMSE

(d) Run Time

Figure 6: ISLET vs. non-convex PGD, Tucker regression, convex regularization. Here, $\sigma = 5$; Panels (a)(b): $p = 10$; Panels (c)(d): $p = 50$.

Next, we investigate the performance of ISLET when p and n substantially grow. Let $p = 100, 150, 200$, $r = 3, 5$, $n \in [8000, 20000]$. The results in RMSE and run time are shown in Fig. 7 (a), (b), (c), and (d), respectively. We can see that the estimation error significantly decays as the sample size n grows, the dimension p decreases, or the Tucker rank r decreases.

We further fix $r = 2, n = 30000$ and let p grow to 400. Now the space cost for storing $\{\boldsymbol{\mathcal{X}}_i\}_{i=1}^n$ reaches $400^3 \times 30000 \times 4$bytes $= 7.68$ terabytes, which is far beyond the volume of most personal computing devices. Since each sample is used only twice in ISLET, we perform this experiment in a parallel way. To be specific, in each machine $b = 1, \ldots, 40$, we store the random seed, draw pseudo random tensor $\boldsymbol{\mathcal{X}}_{bi}$, evaluate y_{bi} and $\widetilde{\boldsymbol{\mathcal{A}}}_b$ by the procedure in Section 2.2, and clean up the memory of $\boldsymbol{\mathcal{X}}_{bi}$. After synchronizing the outcomes and obtaining the importance sketching directions, for each machine $b = 1, \ldots, 40$, we generate pseudo-random covariates $\boldsymbol{\mathcal{X}}_{bi}$ again using the stored random seeds, evaluate $\widetilde{\mathbf{G}}_b$ and $\widetilde{\mathbf{X}}_{bi}$ by (11)-(12), and clean up the memory of $\boldsymbol{\mathcal{X}}_{bi}$ again. The rest of the procedure follows from Section 2.2 and the original ISLET in Algorithm 1. The average RMSE and run time for five repeats are shown in Figure 8. We clearly see that ISLET yields good statistical performance within a reasonable amount of time, while the other contemporary methods can hardly do so in such a ultrahigh-dimensional setting.

In addition, we explore the numerical performance of ISLET for simultaneously sparse and low-rank tensor regression. To perform sparse ISLET (Algorithm 2), we apply the *gglasso* package[3] [131] for group Lasso and penalty level selection. Let n vary from 1500 to 4000, $p = 20, 25, 30$, $r = 3, 5$, $\sigma = 5$, $s = s_1 = s_2 = s_3 = 8$. The result is shown in Fig. 9. Similar to the non-sparse ISLET, as sample size n increases or Tucker rank r decreases, the average estimation errors decrease.

We also compare sparse ISLET with slice-sparse non-convex PGD proposed by [29]. Let $n \in [5000, 12000]$, $p = 50$, $r = 3, 5$, $\sigma = 5$, $s_1 = s_2 = s_3 = 15$. From Fig. 10, we can see that ISLET yields much smaller estimation error with significantly shorter time than non-convex PGD – the difference between two algorithms becomes more significant as n grows.

Finally, if the tensor is of order 2, tensor regression becomes the classic *low-rank matrix recovery* problem [25, 104]. Among existing approaches for low-rank matrix recovery, the nuclear norm minimization (NNM) has been proposed and extensively studied in recent literature. We compare the numerical performance of matrix ISLET (see Algorithm 4 in Section C for implementation details) and NNM that aims to solve [4]

$$\sum_{i=1}^{n}(y_i - \langle \mathbf{X}_i, \mathbf{A}\rangle)^2 + \lambda||\mathbf{A}||_*,$$

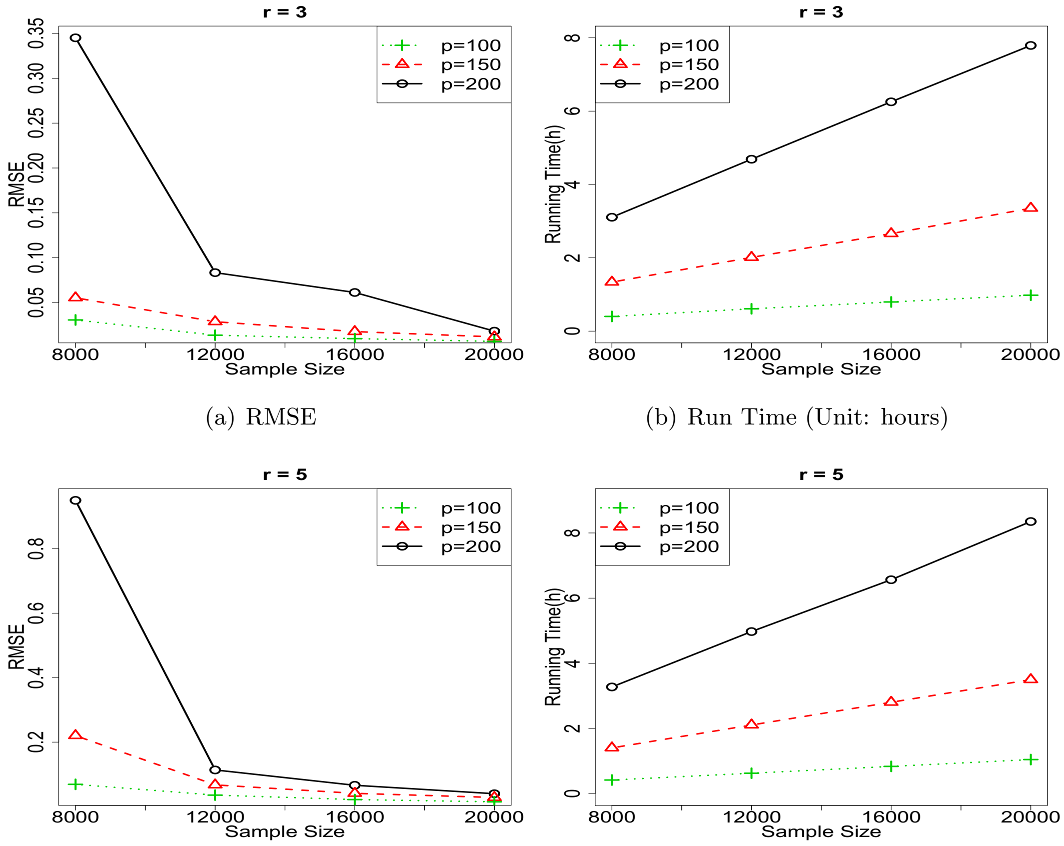
where $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A})$ is the matrix nuclear norm. We consider two specific settings: (1) $p_1 = p_2 = 50$, $r = 2$, $\sigma = 10$, $n \in [2000, 16000]$; (2) $p_1 = p_2 = 100, r = 4, \sigma = 10, n \in [2000, 28000]$. From Figure 11, we find that ISLET has similar, or sometimes even better performance than NNM in estimation error. On the other hand, the run time of ISLET is negligibly small compared to NNM.

6 Discussion

In this article, we develop a general importance sketching algorithm for high-dimensional low-rank tensor regression. In particular, to sufficiently reduce the dimension of the higher-

[3] Available online at: `https://cran.r-project.org/web/packages/gglasso/index.html`.

[4] The optimization of NNM is implemented by accelerated proximal gradient method [115] using the software package available online at `http://www.math.nus.edu.sg/~mattohkc/NNLS.html`.



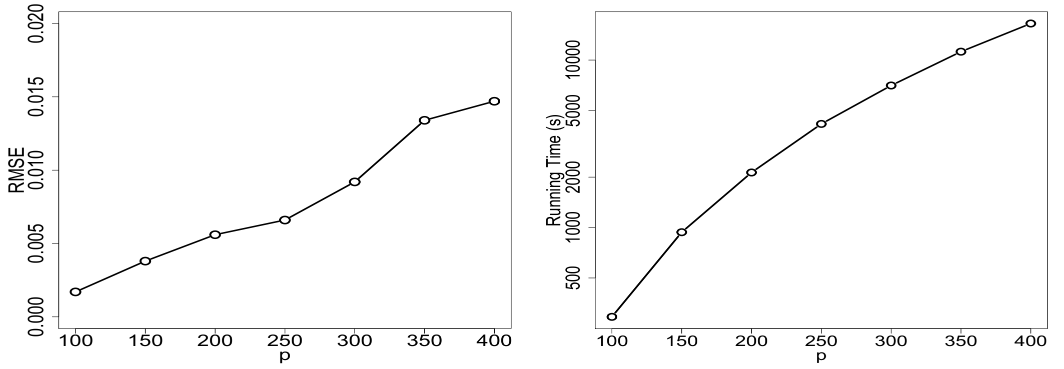
(a) RMSE

(b) Run Time (Unit: hours)

(c) RMSE

(d) Run Time (Unit: hours)

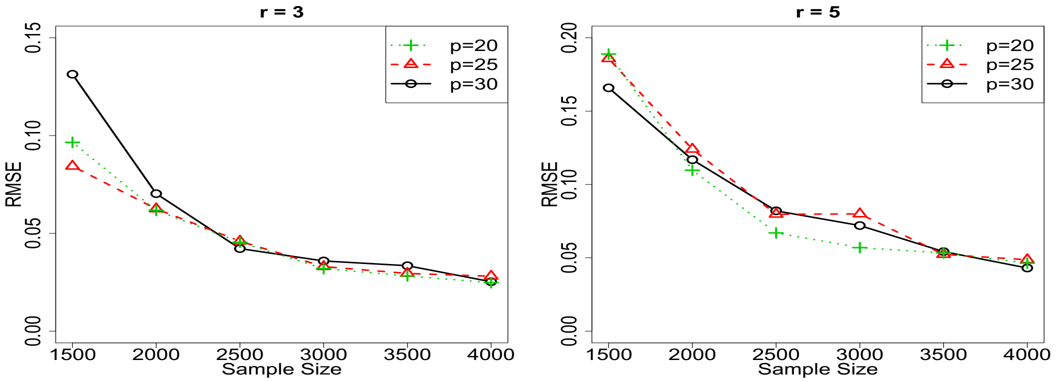
Figure 7: Performance of ISLET when p and n significantly grow.



(a) RMSE

(b) Run Time

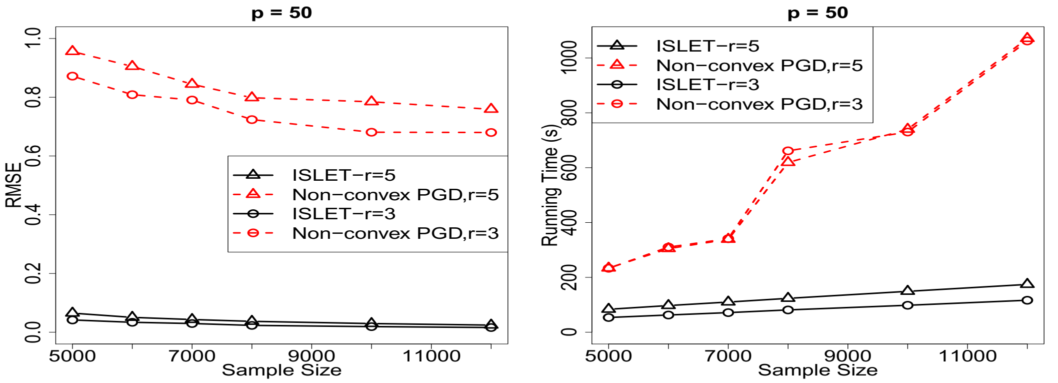
Figure 8: Performance of ISLET in ultrahigh-dimensional setting. p grows up to 400, $n = 30000$.



(a) $r = 3$

(b) $r = 5$

Figure 9: RMSE of ISLET for sparse and low-rank tensor recovery

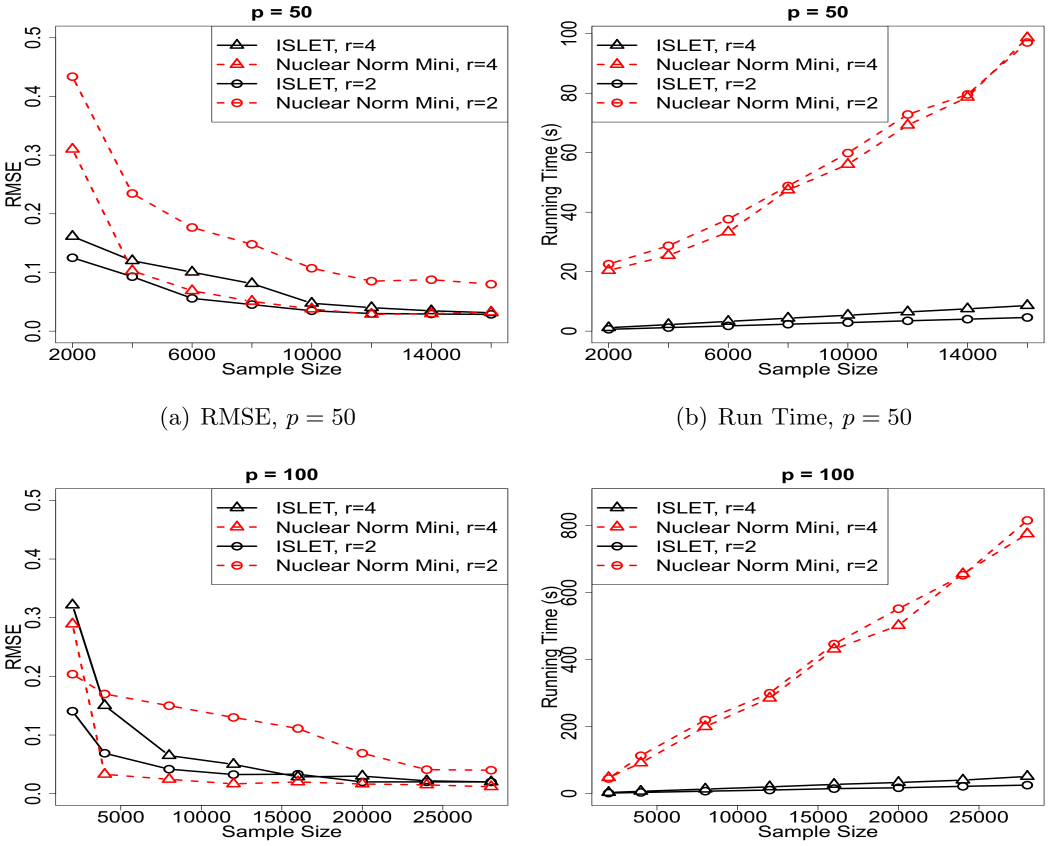


(a) RMSE

(b) Run Time

Figure 10: ISLET vs. non-convex PGD for sparse tensor regression

order structure, we propose a fast algorithm named *importance sketching low-rank estimation for tensors* (ISLET). The proposed algorithm includes three major steps: we first apply tensor decomposition approaches, such as HOOI and STAT-SVD, to obtain importance sketching directions; then we perform regression using the sketched tensor/matrices (in the sparse case, we add group-sparsity regularizers); finally we assemble the final estimator. We establish deterministic oracle inequalities for the proposed procedure under general design and noise distributions. We also prove that ISLET achieves optimal mean-squared error rate under Gaussian ensemble design – regular ISLET can further achieves the optimal constant for mean-squared error. As illustrated in simulation studies, the proposed procedure is computationally efficient comparing to contemporary methods. Although the presentation mainly focuses on order-3 tensors here, the method and theory for the general order-d tensors can be elaborated similarly.



(a) RMSE, $p = 50$

(b) Run Time, $p = 50$

(c) RMSE, $p = 100$

(d) Run Time, $p = 100$

Figure 11: ISLET vs. nuclear norm minimization for low-rank matrix recovery

It is also noteworthy that the storage cost for Tucker decomposition in the proposed procedure grows exponentially with the order d. Thus, if the target tensor has a large order, it is more desirable to consider other low-rank approximation methods than Tucker, such as the CP decomposition [12, 13], Hierarchical Tucker (HT) decomposition [7, 50, 54], and Tensor Train (TT) decomposition [93, 96], etc. The ISLET framework can be adapted to these structures as long as there are two key components: there exists a sketching approach for dimension reduction and a computational inversion step for embedding the low-dimensional estimate back to the high-dimensional space (also see Section 2.4). Whether these components hold for the previously described methods remains an interesting open question.

In addition to low-rank tensor regression, the idea of ISLET can be applied to various other high-dimensional problems. First, *high-order interaction pursuit* is an important topic in high-dimensional statistics that aims at the interaction among three or more variables in the regression setting. This problem can be transformed to the tensor estimation based on a number of rank-1 projections by the argument in [55]. Similarly to analysis on tensor regression in this paper, the idea of ISLET can be used to develop an optimal and efficient procedure for high-order interaction pursuit with provable advantages over other baseline methods.

In addition, *matrix/tensor completion* has attracted significant attention in the recent literature [27, 78, 127, 128, 134]. The central task of matrix/tensor completion is to complete the low-rank matrix/tensor based on a limited number of observable entries. Since each observable entry in matrix/tensor completion can be seen as a special rank-one projection of the original matrix/tensor, the idea behind ISLET can be used to achieve a more efficient algorithm in matrix/tensor completion with theoretical guarantees. It will be an interesting future topic to further investigate the performance of ISLET on other high-dimensional problems.

References

[1] Genevera I Allen. Regularized tensor factorizations and higher-order principal components analysis. *arXiv preprint arXiv:1202.2476*, 2012.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[3] Haim Avron, Kenneth L Clarkson, and David P Woodruff. Sharper bounds for regression and low-rank approximation with regularization. *arXiv preprint arXiv:1611.03225*, 6, 2016.

[4] Haim Avron, Huy Nguyen, and David Woodruff. Subspace embeddings for the polynomial kernel. In *Advances in Neural Information Processing Systems*, pages 2258–2266, 2014.

[5] Krishnakumar Balasubramanian, Jianqing Fan, and Zhuoran Yang. Tensor methods for additive index models under discordance and heterogeneity. *arXiv preprint arXiv:1807.06693*, 2018.

[6] Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression. *arXiv preprint arXiv:1803.07054*, 2018.

[7] Jonas Ballani and Lars Grasedyck. A projection method to solve linear systems in tensor format. *Numerical linear algebra with applications*, 20(1):27–43, 2013.

[8] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P Woodruff. A ptas for ℓ_p-low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 747–766. SIAM, 2019.

[9] Mario Bebendorf. Adaptive cross approximation of multivariate functions. *Constructive approximation*, 34(2):149–179, 2011.

[10] Gregory Beylkin and Martin J Mohlenkamp. Algorithms for numerical analysis in high dimensions. *SIAM Journal on Scientific Computing*, 26(6):2133–2159, 2005.

[11] Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics*, 46(6B):3308–3333, 2018.

[12] M Boussé, I Domanov, and L De Lathauwer. Linear systems with a multilinear singular value decomposition constrained solution. *ESAT-STADIUS, KU Leuven, Belgium, Tech. Rep*, 2017.

[13] Martijn Boussé, Nico Vervliet, Ignat Domanov, Otto Debals, and Lieven De Lathauwer. Linear systems with a canonical polyadic decomposition constrained solution: Algorithms and applications. *Numerical Linear Algebra with Applications*, 25(6):e2190, 2018.

[14] Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. *SIAM Journal on Computing*, 46(2):543–589, 2017.

[15] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

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

[17] T Tony Cai and Anru Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE transactions on information theory*, 60(1):122–132, 2014.

[18] T Tony Cai and Anru Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[19] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

[20] Tianxi Cai, T. Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514):621–633, 2016.

[21] Cesar F Caiafa and Andrzej Cichocki. Generalizing the column–row matrix decomposition to multi-way arrays. *Linear Algebra and its Applications*, 433(3):557–573, 2010.

[22] Raffaello Camoriano, Tomás Angles, Alessandro Rudi, and Lorenzo Rosasco. Nytro: When subsampling meets early stopping. In *Artificial Intelligence and Statistics*, pages 1403–1411, 2016.

[23] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[24] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[25] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[26] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[27] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[28] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[29] Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *arXiv preprint arXiv:1611.10349*, 2016.

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

[31] Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P Woodruff. Algorithms for ℓ_p low-rank approximation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 806–814. JMLR. org, 2017.

[32] Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.

[33] Kenneth L Clarkson and David P Woodruff. Input sparsity and hardness for robust subspace approximation. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 310–329. IEEE, 2015.

[34] Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54, 2017.

[35] Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Robert D Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

[36] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[37] Huaian Diao, Zhao Song, Wen Sun, and David Woodruff. Sketching for kronecker product regression and p-splines. In *International Conference on Artificial Intelligence and Statistics*, pages 1299–1308, 2018.

[38] Edgar Dobriban and Sifan Liu. A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*, 2018.

[39] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

[40] Petros Drineas and Michael W Mahoney. Effective resistances, statistical leverage, and applications to linear equation solving. *arXiv preprint arXiv:1005.3097*, 2010.

[41] Lars Eldén and Berkant Savas. A newton–grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. *SIAM Journal on Matrix Analysis and applications*, 31(2):248–271, 2009.

[42] Mike Espig, Wolfgang Hackbusch, Thorsten Rohwedder, and Reinhold Schneider. Variational calculus with sums of elementary tensors of fixed rank. *Numerische Mathematik*, 122(3):469–488, 2012.

[43] Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *arXiv preprint arXiv:1710.08083*, 2017.

[44] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[45] Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

[46] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[47] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[48] Irina Georgieva and Clemens Hofreither. Greedy low-rank approximation in tucker format of solutions of tensor linear systems. *Journal of Computational and Applied Mathematics*, 358:206–220, 2019.

[49] SA Goreinov, Ivan V Oseledets, and Dmitry V Savostyanov. Wedderburn rank reduction and krylov subspace method for tensor approximation. part 1: Tucker case. *SIAM Journal on Scientific Computing*, 34(1):A1–A27, 2012.

[50] Lars Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2029–2054, 2010.

[51] Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

[52] Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian tensor regression. *arXiv preprint arXiv:1509.06490*, 2015.

[53] Weiwei Guo, Irene Kotsia, and Ioannis Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2012.

[54] Wolfgang Hackbusch and Stefan Kühn. A new scheme for the tensor representation. *Journal of Fourier analysis and applications*, 15(5):706–722, 2009.

[55] Botao Hao, Anru Zhang, and Guang Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *arXiv preprint arXiv:1801.09326*, 2018.

[56] Jarvis Haupt, Xingguo Li, and David P Woodruff. Near optimal sketching of low-rank tensor regression. *arXiv preprint arXiv:1709.07093*, 2017.

[57] Shiyuan He, Jianxin Yin, Hongzhe Li, and Xing Wang. Graphical model selection and estimation for high dimensional tensor data. *Journal of Multivariate Analysis*, 128:165–185, 2014.

[58] Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3):1169, 2015.

[59] Clemens Hofreither. A black-box low-rank approximation algorithm for fast matrix assembly in isogeometric analysis. *Computer Methods in Applied Mechanics and Engineering*, 333:311–330, 2018.

[60] Thomas JR Hughes, John A Cottrell, and Yuri Bazilevs. Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer methods in applied mechanics and engineering*, 194(39-41):4135–4195, 2005.

[61] Mariya Ishteva, P-A Absil, Sabine Van Huffel, and Lieven De Lathauwer. Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

[62] Mariya Ishteva, Lieven De Lathauwer, P-A Absil, and Sabine Van Huffel. Differential-geometric newton method for the best rank-(r 1, r 2, r 3) approximation of tensors. *Numerical Algorithms*, 51(2):179–194, 2009.

[63] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

[64] Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

[65] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[66] Tamara Gibson Kolda. *Multilinear operators for higher-order decompositions*, volume 2. United States. Department of Energy, 2006.

[67] Vladimir Koltchinskii. A remark on low rank matrix recovery and noncommutative bernstein type inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 213–226. Institute of Mathematical Statistics, 2013.

[68] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[69] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Preconditioned low-rank riemannian optimization for linear systems with tensor product structure. *SIAM Journal on Scientific Computing*, 38(4):A2018–A2044, 2016.

[70] Daniel Kressner and Christine Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM journal on matrix analysis and applications*, 31(4):1688–1714, 2010.

[71] Pieter M Kroonenberg. *Applied multiway data analysis*, volume 702. John Wiley & Sons, 2008.

[72] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[73] Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Ruslan R Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[74] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[75] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, pages 1–16, 2017.

[76] Nan Li and Baoxin Li. Tensor completion for on-board compression of hyperspectral images. In *2010 IEEE International Conference on Image Processing*, pages 517–520. IEEE, 2010.

[77] Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, pages 1–26, 2018.

[78] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[79] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[80] RE Lynch, JOHN R Rice, and DONALD H Thomas. Tensor product analysis of partial difference equations. *Bulletin of the American Mathematical Society*, 70(3):378–384, 1964.

[81] Xiang Lyu, Will Wei Sun, Zhaoran Wang, Han Liu, Jian Yang, and Guang Cheng. Tensor graphical model: Non-convex optimization and statistical inference. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[82] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[83] Michael W Mahoney, Mauro Maggioni, and Petros Drineas. Tensor-cur decompositions for tensor-based data. *SIAM Journal on Matrix Analysis and Applications*, 30(3):957–987, 2008.

[84] Ameur M Manceur and Pierre Dutilleul. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49, 2013.

[85] Panos P Markopoulos, George N Karystinos, and Dimitris A Pados. Optimal algorithms for l_{1}-subspace signal processing. *IEEE Transactions on Signal Processing*, 62(19):5046–5058, 2014.

[86] Panos P Markopoulos, Sandipan Kundu, Shubham Chamadia, and Dimitris A Pados. Efficient l1-norm principal-component analysis via bit flipping. *IEEE Transactions on Signal Processing*, 65(16):4252–4264, 2017.

[87] Pascal Massart. *Concentration inequalities and model selection*. Springer, 2007.

[88] Deyu Meng, Zongben Xu, Lei Zhang, and Ji Zhao. A cyclic weighted median method for l1 low-rank matrix factorization with missing entries. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[89] Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100. ACM, 2013.

[90] Andrea Montanari and Nike Sun. Spectral algorithms for tensor completion. *arXiv preprint arXiv:1612.07866*, 2016.

[91] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *ICML*, pages 73–81, 2014.

[92] Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 117–126. IEEE, 2013.

[93] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[94] Ivan V Oseledets, DV Savostianov, and Eugene E Tyrtyshnikov. Tucker dimensionality reduction of three-dimensional arrays in linear time. *SIAM Journal on Matrix Analysis and Applications*, 30(3):939–956, 2008.

[95] Ivan V Oseledets, Dmitry V Savostyanov, and Eugene E Tyrtyshnikov. Cross approximation in tensor electron density computations. *Numerical Linear Algebra with Applications*, 17(6):935–952, 2010.

[96] Ivan V Oseledets and Eugene E Tyrtyshnikov. Breaking the curse of dimensionality, or how to use svd in many dimensions. *SIAM Journal on Scientific Computing*, 31(5):3744–3759, 2009.

[97] Rasmus Pagh. Compressed matrix multiplication. *ACM Transactions on Computation Theory (TOCT)*, 5(3):9, 2013.

[98] Yuqing Pan, Qing Mai, and Xin Zhang. Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association*, pages 1–15, 2018.

[99] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013.

[100] Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

[101] Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

[102] Garvesh Raskutti and Michael Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *arXiv preprint arXiv:1406.5986*, 2014.

[103] Garvesh Raskutti, Ming Yuan, and Han Chen. Convex regularization for high-dimensional multi-response tensor regression. *arXiv preprint arXiv:1512.01215*, 2015.

[104] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[105] Berkant Savas and Lars Eldén. Krylov-type methods for tensor computations i. *Linear Algebra and its Applications*, 438(2):891–918, 2013.

[106] Berkant Savas and Lek-Heng Lim. Quasi-newton methods on grassmannians and multilinear approximations of tensors. *SIAM Journal on Scientific Computing*, 32(6):3352–3393, 2010.

[107] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

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

[109] Nicholas D Sidiropoulos, Evangelos E Papalexakis, and Christos Faloutsos. Parallel randomly compressed cubes: A scalable distributed architecture for big tensor decomposition. *IEEE Signal Processing Magazine*, 31(5):57–70, 2014.

[110] Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l 1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701. ACM, 2017.

[111] Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2772–2789. Society for Industrial and Applied Mathematics, 2019.

[112] Will Wei Sun and Lexin Li. Sparse low-rank tensor response regression. *arXiv preprint arXiv:1609.04523*, 2016.

[113] Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

[114] Yiming Sun, Yang Guo, Charlene Luo, Joel Tropp, and Madeleine Udell. Low-rank tucker approximation of a tensor from streaming data. *arXiv preprint arXiv:1904.10951*, 2019.

[115] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[116] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *Advances in neural information processing systems*, pages 1331–1339, 2013.

[117] Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems*, pages 972–980, 2011.

[118] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017.

[119] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973, 2016.

[120] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[121] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[122] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[123] Nico Vervliet and Lieven De Lathauwer. A randomized block sampling approach to canonical polyadic decomposition of large-scale tensors. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):284–295, 2015.

[124] Jialei Wang, Jason D Lee, Mehrdad Mahdavi, Mladen Kolar, Nathan Srebro, et al. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2):4896–4944, 2017.

[125] Yining Wang, Hsiao-Yu Tung, Alexander J Smola, and Anima Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *Advances in Neural Information Processing Systems*, pages 991–999, 2015.

[126] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[127] Dong Xia and Ming Yuan. On polynomial time methods for exact low rank tensor completion. *arXiv preprint arXiv:1702.06980*, 2017.

[128] Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

[129] Lingzhou Xue and Hui Zou. Sure independence screening and compressed random sensing. *Biometrika*, pages 371–380, 2011.

[130] Dan Yang, Zongming Ma, and Andreas Buja. A sparse singular value decomposition method for high-dimensional data. *Journal of Computational and Graphical Statistics*, 23(4):923–942, 2014.

[131] Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

[132] Ming Yu, Zhaoran Wang, Varun Gupta, and Mladen Kolar. Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach. *arXiv preprint arXiv:1802.06967*, 2018.

[133] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[134] Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, pages 1–38, 2014.

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

[136] Anru Zhang and Rungang Han. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, page to appear, 2018.

[137] Anru Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. Supplement to “ISLET: Fast and optimal low-rank tensor regression via importance sketching”, 2018.

[138] Anru Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. A sharp blockwise tensor perturbation bound for higher-order orthogonal iteration. *preprint*, 2019.

[139] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

[140] Lijun Zhang, Mehrdad Mahdavi, Rong Jin, Tianbao Yang, and Shenghuo Zhu. Random projections for classification: A recovery approach. *IEEE Transactions on Information Theory*, 60(11):7300–7316, 2014.

[141] Yinqiang Zheng, Guangcan Liu, Shigeki Sugimoto, Shuicheng Yan, and Masatoshi Okutomi. Practical low-rank matrix approximation under robust l 1-norm. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1410–1417. IEEE, 2012.

[142] Hua Zhou. Matlab tensorreg toolbox version 1.0, 2017. Available online at https://hua-zhou.github.io/TensorReg/.

[143] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[144] Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

Supplement to “ISLET: Fast and Optimal Low-rank Tensor Regression via Importance Sketching”

Anru Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan

Abstract

In this supplement, we provide additional notation, preliminaries, ISLET procedure for general order tensor estimations, more details on tuning parameter selection, and all proofs for the main results of the paper.

A Additional Notation and Preliminaries

To conveniently specify the dimensions of tensors, for an order-d tensor $\boldsymbol{\mathcal{A}}$ with dimensions $p_1 \times \cdots \times p_d$, we denote $p_{-k} = p_1 \cdots p_d / p_k$ for $k = 1, \ldots, d$. Then the mode-k matricization of $\boldsymbol{\mathcal{A}}$, denoted as $\mathcal{M}_k(\boldsymbol{\mathcal{A}})$, has dimension $p_k \times p_{-k}$. For any matrix $\mathbf{D} \in \mathbb{R}^{p_1 \times p_2}$ and order-d tensor $\boldsymbol{\mathcal{A}}$, we formally define the vectorization as

$$\begin{aligned}
&\mathrm{vec}(\mathbf{D}) \in \mathbb{R}^{(p_1 p_2)}, \quad \mathrm{vec}(\mathbf{D})_{[i_1+(i_2-1)p_1]} = \mathbf{D}_{[i_1,i_2]};\\
&\mathrm{vec}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{(p_1 \cdots p_d)}, \quad \mathrm{vec}(\boldsymbol{\mathcal{A}})_{[i_1+p_1(i_2-1)+\cdots+(i_d-1)p_1\cdots p_d]} = \boldsymbol{\mathcal{A}}_{[i_1,\ldots,i_d]}.
\end{aligned}$$

For any tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, the Mode-$k$ matricization is formally defined as

$$\mathcal{M}_k(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{p_k \times p_{-k}}, \quad \boldsymbol{\mathcal{A}}_{[i_1,\ldots,i_d]} = \left(\mathcal{M}_k(\boldsymbol{\mathcal{A}})\right)_{[i_k,j]}, \quad j = 1 + \sum_{\substack{l=1\\l\neq k}}^{d} \left\{ (i_l - 1) \prod_{\substack{m=1\\m\neq k}}^{l-1} p_m \right\}$$

for any $1 \leq i_l \leq p_l, l = 1, \ldots, d$. Also see [65, Section 2.4] for more discussions on tensor matricizations.

In order to better illustrate the proposed procedure, we have introduced a row-permutation operator $\mathcal{R}_k$ that matches the index of $\mathbf{W}_k \otimes \mathbf{V}_k$ to $\mathrm{vec}(\boldsymbol{\mathcal{A}})$. In particular if $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, $\mathbf{W}_k \in \mathbb{R}^{p_{-k} \times r_k}$, $\mathbf{V}_k \in \mathbb{R}^{p_k \times r_k}$, $\mathcal{R}_k$ is defined as follows:

$$\begin{aligned}
&\left(\mathcal{R}_1\left(\mathbf{W}_1 \otimes \mathbf{V}_1\right)\right)_{[i_1+(i_2-1)p_1+(i_3-1)p_1p_2,:]} = \left(\mathbf{W}_1 \otimes \mathbf{V}_1\right)_{[i_1+(i_2-1)p_1+(i_3-1)p_1p_2,:]},\\
&\left(\mathcal{R}_2\left(\mathbf{W}_2 \otimes \mathbf{V}_2\right)\right)_{[i_1+(i_2-1)p_1+(i_3-1)p_1p_2,:]} = \left(\mathbf{W}_2 \otimes \mathbf{V}_2\right)_{[i_2+(i_1-1)p_2+(i_3-1)p_2p_1,:]},\\
&\left(\mathcal{R}_3\left(\mathbf{W}_3 \otimes \mathbf{V}_3\right)\right)_{[i_1+(i_2-1)p_1+(i_3-1)p_1p_2,:]} = \left(\mathbf{W}_3 \otimes \mathbf{V}_3\right)_{[i_3+(i_1-1)p_3+(i_2-1)p_1p_3,:]}
\end{aligned}$$

for $1 \leq i_1 \leq p_1, 1 \leq i_2 \leq p_2, 1 \leq i_3 \leq p_3$.

B ADHD MRI Imaging Data Analysis

In this section, we display the value of our method on predicting attention deficit hyperactivity disorder (ADHD) with magnetic resonance imaging (MRI) dataset provided by Neuro Bureau[5]. The dataset involves 973 subjects, where each subject is associated with a 121-by-145-by-121 MRI image and several demographic variables. After removing the missing values, we obtain 930 samples, among which 356 and 574 are diagnosed and control subjects, respectively.

We aim to do prediction based on the association between the diagnosis label y_i of i^{th} observation and its covariates with MRI imaging $\boldsymbol{\mathcal{X}}_i$, demographic variables age x_i^1, gender x_i^2, and handedness x_i^3. To better cope the job of predicting binary response y_i and incorporate the demographic information in addition to tensor image covariates, we apply importance sketching, the central idea of ISLET, for dimension reduction. The 5-fold cross-validation is applied to examine the prediction power. Specifically for $l = 1, \dots, 50$, we randomly partition all 930 subjects into 5 uniform subsets $\{\Omega_j^{(l)}\}_{j=1,\dots,5} \subseteq \{1, \dots, 930\}$. For $j = 1, \dots, 5$, we assign one fold $\Omega_j^{(l)}$ and the other four folds $\Omega_{-j}^{(l)} = \cup_{j' \neq j} \Omega_{j'}$ as the testing and training sets, respectively. We apply Step 1 of sparse ISLET (described in Section 2.3) on $\{y_i, \boldsymbol{\mathcal{X}}_i\}_{i \in \Omega_{(-j)}^{(l)}}$ to obtain $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3$ and construct the importance sketching covariates $\widetilde{\mathbf{x}}_i = \text{vec}(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{U}_1^\top, \times_1 \widetilde{U}_1^\top, \times_1 \widetilde{U}_1^\top)$, perform logistic regression for y_i versus the combined covariates $\left[\widetilde{\mathbf{x}}_i, x_i^1, x_i^2, x_i^3\right]$, $i \in \Omega_{-j}^{(l)}$ and possible ℓ_1 regularizer to get the estimates. Then we use estimates and $\left[\widetilde{\mathbf{x}}_i, x_1^i, x_2^i, x_3^i\right], i \in \Omega_j^{(l)}$ to predict the labels of samples in the testing set $\Omega_j^{(l)}$. For comparison, we also perform Tucker regression and Tucker regression with regularizer proposed by [77, 143] under the same setting. Since it is computationally intensive to perform full Tucker regression on complete tensor covariates of dimension $121 \times 145 \times 121$, we follow the procedure described in [77, 143] and apply the discrete cosine transformation to downsize the MRI data to $12 \times 14 \times 12$ using the code available at the authors' website [142]. For all methods, we input Tucker rank (r, r, r) for $r = 3, 4, 5$ and other regularization tuning parameters selected via cross validation. We repeat experiments for $l = 1, \dots, 50, j = 1, \dots, 5$ and take average to ensure stable estimations of the prediction accuracy for both procedures.

The average prediction accuracy with standard deviation in the parenthesis and runtime for both methods are shown in Table 1. We can see the importance sketching method performs significantly better than Tucker regression in both the prediction accuracy and runtime for all different Tucker rank choices. Particularly for the importance sketching, adding ℓ_1 regularizer provides more accurate prediction but costs more time. In addition, compared to the downsizing method by [143, 77] that deterministically relies on external

[5] Link: `http://neurobureau.projects.nitrc.org/ADHD200/Data.html`

	Rank	Methods			
		IS	IS	Tucker Reg.	Tucker Reg.
			+ regularizer		+ regularizer
Prediction	3	0.684(0.010)	**0.686**(0.009)	0.624(0.014)	0.647(0.009)
Accuracy	4	0.673(0.009)	**0.682**(0.008)	0.609(0.014)	0.648(0.007)
	5	0.653(0.009)	**0.674**(0.007)	0.591(0.015)	0.644(0.007)
Runtime	3	**0.008**	0.392	14.291	3.03
Unit:	4	**0.024**	1.003	22.088	5.761
seconds	5	**0.064**	3.339	33.392	13.710

Table 1: Importance sketching (IS) vs. Tucker regression in prediction accuracy and runtime

information, our importance sketching is fully data-driven. We can also see downsizing the tensor covariates to 3-by-3-by-3 by importance sketching provides more prediction power than downsizing to 12-by-14-by-12 by deterministic methods. This reveals the runtime advantage and immediately demonstrates the advantage of the proposed method over other state-of-the-art approaches.

C ISLET for General Order Tensor Estimation

For completeness, we provide the ISLET procedure for general order-d low-rank tensor estimation in this section. The procedure for $d \geq 3$ is provided in Algorithms 3 and the one for $d = 2$ (i.e., the low-rank matrix estimation) is provided in Algorithm 4. The sparse versions for $d \geq 3$ and $d = 2$ are provided in Algorithms 5 and 6, respectively.

D More Details on Tuning Parameter Selection

The implementation of ISLET requires the rank $\boldsymbol{r}$ as inputs. When $\boldsymbol{r}$ is unknown in practice, we propose a two-stage-scheme for adaptive low-rank tensor regression. First, we input a conservatively large value of $\boldsymbol{r}_{ini}$ into ISLET to obtain $\widehat{\boldsymbol{\mathcal{B}}}, \widehat{\mathbf{D}}_k$ (regular case) or $\widehat{\boldsymbol{\mathcal{B}}}, \widehat{\mathbf{E}}_k$ (sparse case), based on which we estimate the rank $\widehat{\boldsymbol{r}}$ by the "Cross scheme" introduced recently by [135]. Then, we run ISLET again with $\widehat{\boldsymbol{r}}$ to obtain the final estimates. The pseudo-codes for regular and sparse order-d tensor regression are provided in Algorithms 7 and 8, respectively.

Next, we perform simulation studies to verify the proposed rank selection scheme in both the regular and sparse cases. In particular, let $p = 20, 30$, $\boldsymbol{r}_{ini} = \lfloor \boldsymbol{p}/3 \rfloor$, $n \in [2000, 5000]$, $\sigma = 5$, $s = 12$, and the actual rank $r = 3, 5$. We randomly generate the regular and sparse regression settings as described in Section 5, then perform Algorithms 7 and 8. The average

Algorithm 3 Order-d ISLET ($d \geq 3$)

```
Input: $y_1, \ldots, y_n \in \mathbb{R}, \boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, rank $\boldsymbol{r} = (r_1, \ldots, r_d)$.
Evaluate $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n} \sum_{j=1}^{n} y_j \boldsymbol{\mathcal{X}}_j$.
Apply order-$d$ HOOI on $\widetilde{\boldsymbol{\mathcal{A}}}$ to obtain initial estimates $\widetilde{\mathbf{U}}_k, k = 1, \ldots, d$.
Let $\widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \ldots, \widetilde{\mathbf{U}}_d^\top]\!]$. Evaluate the sketching directions,
    $\widetilde{\mathbf{V}}_k = \mathrm{QR}\left[\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right], \quad k = 1, \ldots, d.$
Construct $\widetilde{\mathbf{X}} = \left[\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \; \widetilde{\mathbf{X}}_{\mathbf{D}_1} \; \cdots \; \widetilde{\mathbf{X}}_{\mathbf{D}_d}\right] \in \mathbb{R}^{n \times m}$, where
    $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n \times m_{\boldsymbol{\mathcal{B}}}}, \quad (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_{l=1}^{d} \widetilde{\mathbf{U}}_l^\top\right),$
    $\widetilde{\mathbf{X}}_{\mathbf{D}_k} \in \mathbb{R}^{n \times m_{\mathbf{D}_k}}, \quad (\widetilde{\mathbf{X}}_{\mathbf{D}_k})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{\substack{l=1\\l\neq k}}^{d} \widetilde{\mathbf{U}}_l^\top\right) \widetilde{\mathbf{V}}_k\right)$
    for $m_{\boldsymbol{\mathcal{B}}} = r_1 \cdots r_d, m_{\mathbf{D}_k} = (p_k - r_k) r_k$, $k = 1, \ldots, d$, and $m = m_{\boldsymbol{\mathcal{B}}} + m_{\mathbf{D}_1} + \cdots + m_{\mathbf{D}_d}$.
Solve $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^m} \|y - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|_2^2$. Partition $\widehat{\boldsymbol{\gamma}}$ to $\widehat{\boldsymbol{\mathcal{B}}}, \widehat{\mathbf{D}}_1, \ldots, \widehat{\mathbf{D}}_d$,
    $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) := \widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} = \widehat{\boldsymbol{\gamma}}_{[1:m_{\boldsymbol{\mathcal{B}}}]},$
    $\mathrm{vec}(\widehat{\mathbf{D}}_k) := \widehat{\boldsymbol{\gamma}}_{\mathbf{D}_k} = \widehat{\boldsymbol{\gamma}}_{\left[\left(m_{\boldsymbol{\mathcal{B}}} + \sum_{k'=1}^{k-1} m_{\mathbf{D}_{k'}} + 1\right):\left(m_{\boldsymbol{\mathcal{B}}} + \sum_{k'=1}^{k} m_{\mathbf{D}_{k'}}\right)\right]}, \quad k = 1, \ldots, d.$
Let $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$, evaluate
    $\widehat{\boldsymbol{\mathcal{A}}} = [\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{L}}_1, \ldots, \widehat{\mathbf{L}}_d]\!], \quad \widehat{\mathbf{L}}_k = \left(\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widehat{\mathbf{D}}_k\right)\left(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right)^{-1}, \quad k = 1, \ldots, d.$
```

Algorithm 4 Matrix ISLET

```
Input: $y_1, \dots, y_n \in \mathbb{R}, \mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^{p_1 \times p_2}$, rank $r$.
Evaluate $\widetilde{\mathbf{A}} = \frac{1}{n}\sum_{j=1}^{n} y_j \mathbf{X}_j$. and let $\widetilde{\mathbf{U}}_1 = \mathrm{SVD}_r(\widetilde{\mathbf{A}}), \widetilde{\mathbf{U}}_2 = \mathrm{SVD}_r(\widetilde{\mathbf{A}}^\top)$.
Construct $\widetilde{\mathbf{X}} = \left[\widetilde{\mathbf{X}}_{\mathbf{B}}\ \widetilde{\mathbf{X}}_{\mathbf{D}_1} \widetilde{\mathbf{X}}_{\mathbf{D}_2}\right] \in \mathbb{R}^{n \times r(p_1+p_2-r)}$, where
    $$\widetilde{\mathbf{X}}_{\mathbf{B}} \in \mathbb{R}^{n \times r^2}, \quad (\widetilde{\mathbf{X}}_{\mathbf{B}})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_1^\top \mathbf{X}_i \widetilde{\mathbf{U}}_2\right),$$
    $$\widetilde{\mathbf{X}}_{\mathbf{D}_k} \in \mathbb{R}^{n \times (p_k - r)r}, \quad (\widetilde{\mathbf{X}}_{\mathbf{D}_1})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{X}_i \widetilde{\mathbf{U}}_2\right), \quad (\widetilde{\mathbf{X}}_{\mathbf{D}_2})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{2\perp}^\top \mathbf{X}_i^\top \widetilde{\mathbf{U}}_1\right).$$
Solve $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^m} \|y - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\|_2^2$. Partition $\widehat{\boldsymbol{\gamma}}$ and assign to $\widehat{\mathbf{B}}, \widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2$,
    $$\mathrm{vec}(\widehat{\mathbf{B}}) := \widehat{\boldsymbol{\gamma}}_{[1:r^2]}, \quad \mathrm{vec}(\widehat{\mathbf{D}}_1) := \widehat{\boldsymbol{\gamma}}_{[(r^2+1):rp_1]}, \quad \mathrm{vec}(\widehat{\mathbf{D}}_2) := \widehat{\boldsymbol{\gamma}}_{[(rp_1+1):(r(p_1+p_2-r))]}.$$
Evaluate
    $$\widehat{\boldsymbol{\mathcal{A}}} = \widehat{\mathbf{L}}_1 \widehat{\mathbf{B}} \widehat{\mathbf{L}}_2^\top, \quad \widehat{\mathbf{L}}_1 = \left(\widetilde{\mathbf{U}}_1 \widehat{\mathbf{B}} + \widetilde{\mathbf{U}}_{1\perp}\widehat{\mathbf{D}}_1\right)\widehat{\mathbf{B}}^{-1}, \quad \widehat{\mathbf{L}}_2 = \left(\widetilde{\mathbf{U}}_2 \widehat{\mathbf{B}}^\top + \widetilde{\mathbf{U}}_{2\perp}\widehat{\mathbf{D}}_2\right)\left(\widehat{\mathbf{B}}^\top\right)^{-1}.$$
```

Algorithm 5 Order-d Sparse ISLET

```
Input: $y_1, \dots, y_n \in \mathbb{R}$, $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, rank $\boldsymbol{r} = (r_1, r_2, \dots, r_d)$, sparsity index $J_s$.
Evaluate $\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n}\sum_{j=1}^{n} y_j \boldsymbol{\mathcal{X}}_j$.
Apply STAT-SVD on $\widetilde{\boldsymbol{\mathcal{A}}}$ with sparsity index $J_s$. Let the outcome be $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3, \dots, \widetilde{\mathbf{U}}_d$.
Let $\widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \dots, \widetilde{\mathbf{U}}_d^\top]\!]$ and evaluate the probing directions $\widetilde{\mathbf{V}}_k = \mathrm{QR}\left[\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right], k = 1, \dots, d$.
Construct
    $$\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n \times (r_1 \cdots r_d)}, \quad (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_{l=1}^{d} \widetilde{\mathbf{U}}_l^\top\right),$$
    $$\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n \times (p_k r_k)}, \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[i,:]} = \mathrm{vec}\left(\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{\substack{l=1\\ l\neq k}}^{d} \widetilde{\mathbf{U}}_l^\top\right)\widetilde{\mathbf{V}}_k\right), \quad k = 1, \dots, d.$$
Solve
    $$\widehat{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{r_1 \cdots r_d}, \quad \mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{r_1 \cdots r_d}}{\arg\min} \|y - \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\boldsymbol{\gamma}\|_2^2;$$
    $$\widehat{\mathbf{E}}_k \in \mathbb{R}^{p_k \times r_k}, \quad \mathrm{vec}(\widehat{\mathbf{E}}_k) = \begin{cases} \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p_k r_k}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2 + \lambda_k \sum_{j=1}^{p_k} \|\boldsymbol{\gamma}_{G_j^k}\|_2, & k \in J_s; \\ \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p_k r_k}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2, & k \notin J_s. \end{cases}$$
Evaluate
    $$\widehat{\boldsymbol{\mathcal{A}}} = [\![\widehat{\boldsymbol{\mathcal{B}}}; (\widehat{\mathbf{E}}_1(\widetilde{\mathbf{U}}_1^\top \widehat{\mathbf{E}}_1)^{-1}), \dots, (\widehat{\mathbf{E}}_d(\widetilde{\mathbf{U}}_d^\top \widehat{\mathbf{E}}_d)^{-1})]\!]$$
```

Algorithm 6 Matrix Sparse ISLET

```
Input: $y_1, \ldots, y_n \in \mathbb{R}$, $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{p_1 \times p_2}$, rank $r$, sparsity index $J_s \subseteq \{1, 2\}$.
Evaluate $\widetilde{\mathbf{A}} = \frac{1}{n_1} \sum_{j=1}^{n} y_j \mathbf{X}_j$. Apply sparse matrix SVD (the Two-Way Iterative Thresholding in [130] or the order-2 version of STAT-SVD in [136]) on $\widetilde{\mathbf{A}}$ with sparsity index $J_s$. Let the estimated left and right subspaces be $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2$.
Construct
    $$\begin{aligned} &\widetilde{\mathbf{X}}_{\mathbf{B}} \in \mathbb{R}^{n \times (r^2)}, \quad (\widetilde{\mathbf{X}}_{\mathbf{B}})_{[i,:]} = \mathrm{vec}(\widetilde{\mathbf{U}}_1^\top \mathbf{X}_i \widetilde{\mathbf{U}}_2), \\ &\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n \times (p_k r)}, \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_1})_{[i,:]} = \mathrm{vec}\left(\mathbf{X}_i \widetilde{\mathbf{U}}_2\right), \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_2})_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_1^\top \mathbf{X}_i\right). \end{aligned}$$
Solve $\widehat{\mathbf{B}} \in \mathbb{R}^{r \times r}, \mathrm{vec}(\widehat{\mathbf{B}}) = \arg\min_{\gamma \in \mathbb{R}^{r^2}} \|y - \widetilde{\mathbf{X}}_{\mathbf{B}} \boldsymbol{\gamma}\|_2^2$;
    $$\widehat{\mathbf{E}}_k \in \mathbb{R}^{p_k \times r_k}, \quad \mathrm{vec}(\widehat{\mathbf{E}}_k) = \begin{cases} \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p_k r}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k} \boldsymbol{\gamma}\|_2^2 + \lambda_k \sum_{j=1}^{p_k} \|\boldsymbol{\gamma}_{G_j^k}\|_2, & k \in J_s; \\ \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p_k r}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k} \boldsymbol{\gamma}\|_2^2, & k \notin J_s. \end{cases}$$
Evaluate
    $$\widehat{\mathbf{A}} = \widehat{\mathbf{E}}_1 (\widetilde{\mathbf{U}}_1^\top \widehat{\mathbf{E}}_1)^{-1} \widehat{\mathbf{B}} (\widetilde{\mathbf{U}}_2^\top \widehat{\mathbf{E}}_2)^{-\top} \widehat{\mathbf{E}}_2^\top.$$
```

Algorithm 7 Order-d ISLET, unknown r

```
Input: $y_1, \ldots, y_n \in \mathbb{R}, \boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, rank $\boldsymbol{r}_{ini} = (r_{1,ini}, \ldots, r_{d,ini})$.
Apply Algorithms 1, 3, 4 with rank $\boldsymbol{r}_{ini}$ to obtain $\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k$, $\widehat{\boldsymbol{\mathcal{B}}}$, and $\widehat{\mathbf{D}}_k$ for $k = 1, \ldots, d$.
Denote $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$. Evaluate $\mathbf{U}_k^{(B)}$ and $\mathbf{V}_k^{(A)}$ via SVDs. Then rotate,
    $$\mathbf{U}_k^{(B)} \in \mathbb{O}_{r_{k,ini}}, \text{ as the left singular vectors of } \widehat{\mathbf{B}}_k,$$
    $$\mathbf{V}_k^{(A)} \in \mathbb{O}_{r_{k,ini}}, \text{ as the right singular vectors of } \left(\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widehat{\mathbf{D}}_k\right);$$
    $$\mathbf{A}_k = \left(\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp} \widehat{\mathbf{D}}_k\right) \mathbf{V}_k^{(A)} \in \mathbb{R}^{p_k \times r_{k,ini}},$$
    $$\boldsymbol{J}_k = (\mathbf{U}_k^{(B)})^\top \cdot \left(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right) \cdot \mathbf{V}_k^{(A)} \in \mathbb{R}^{r_{k,ini} \times r_{k,ini}}.$$
for $k = 1, \ldots, d$ do
    for $s = r_{k,ini} : -1 : 1$ do
        if $\boldsymbol{J}_{k,[1:s,1:s]}$ is not singular and $\|\mathbf{A}_{k,[:,1:s]} \boldsymbol{J}_{k,[1:s,1:s]}^{-1}\| \leq 3$ then
            $\widehat{r}_k = s$; break from the loop;
        end if
    end for
    If $\widehat{r}_k$ is still unassigned then $\widehat{r}_k = 0$.
end for
Apply Algorithm 1 again with rank $\widehat{\boldsymbol{r}} = (\widehat{r}_1, \ldots, \widehat{r}_d)$. Let the final output be $\widehat{\boldsymbol{\mathcal{A}}}$.
```

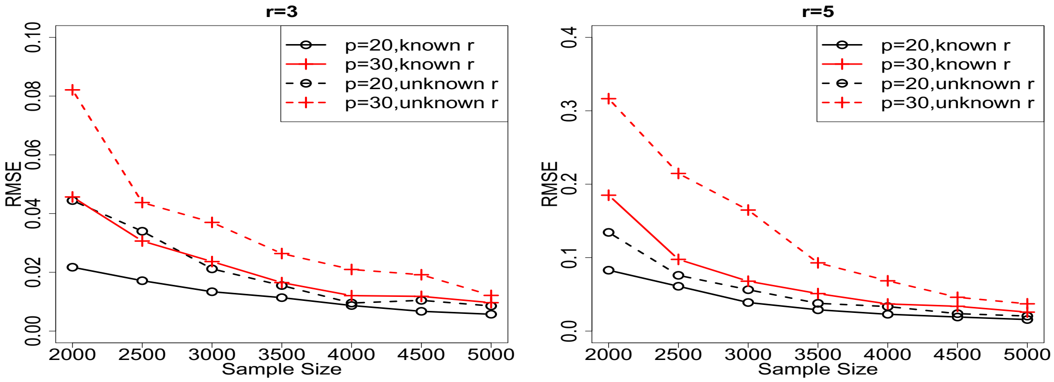
Algorithm 8 Order-d Sparse ISLET, unknown r

```
Input: $y_1, \ldots, y_n \in \mathbb{R}, \boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, rank $\boldsymbol{r}_{ini}$, sparsity index $J_s$.
Apply Algorithms 2, 5, or 6 with rank $\boldsymbol{r}_{ini}$ to obtain $\widetilde{\mathbf{U}}_k$, $\widetilde{\mathbf{V}}_k$, $\widehat{\boldsymbol{\mathcal{B}}}$, and $\widehat{\mathbf{E}}_k$ for $k = 1, \ldots, d$.
Denote $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$. Evaluate $\mathbf{U}_k^{(B)}$ and $\mathbf{V}_k^{(A)}$ via SVDs, then rotate,
    $\mathbf{U}_k^{(B)} \in \mathbb{O}_{r_{k,ini}}$, as the left singular vectors of $\widehat{\mathbf{B}}_k$,
    $\mathbf{V}_k^{(A)} \in \mathbb{O}_{r_{k,ini}}$, as the right singular vectors of $\widehat{\mathbf{E}}_k$;
    $\mathbf{A}_k = \widehat{\mathbf{E}}_k \mathbf{V}_k^{(A)} \in \mathbb{R}^{p_k \times r_{k,ini}}, \quad \boldsymbol{J}_k = (\mathbf{U}_k^{(B)})^\top \cdot \left(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k\right) \cdot \mathbf{V}_k^{(A)} \in \mathbb{R}^{r_{k,ini} \times r_{k,ini}}$.
for $k = 1, \ldots, d$ do
    for $s = r_{k,ini} : -1 : 1$ do
        if $\boldsymbol{J}_{k,[1:s,1:s]}$ is not singular and $\|\mathbf{A}_{k,[:,1:s]} \boldsymbol{J}_{k,[1:s,1:s]}^{-1}\| \leq 3$ then
            $\widehat{r}_k = s$; break from the loop;
        end if
    end for
    If $\widehat{r}_k$ is still unassigned then $\widehat{r}_k = 0$.
end for
Apply Algorithm 2 again with rank $\widehat{\boldsymbol{r}} = (\widehat{r}_1, \ldots, \widehat{r}_d)$. Let the final output be $\widehat{\boldsymbol{\mathcal{A}}}$.
```

estimation error results are plots in Figures 12 and 13 respectively for the regular and sparse cases. We can see from both cases that the estimation errors with known rank are close to the one without known rank and the difference decreases when the sample size gets larger.

E Simulation Study on Approximate Low-rank Tensor Regression

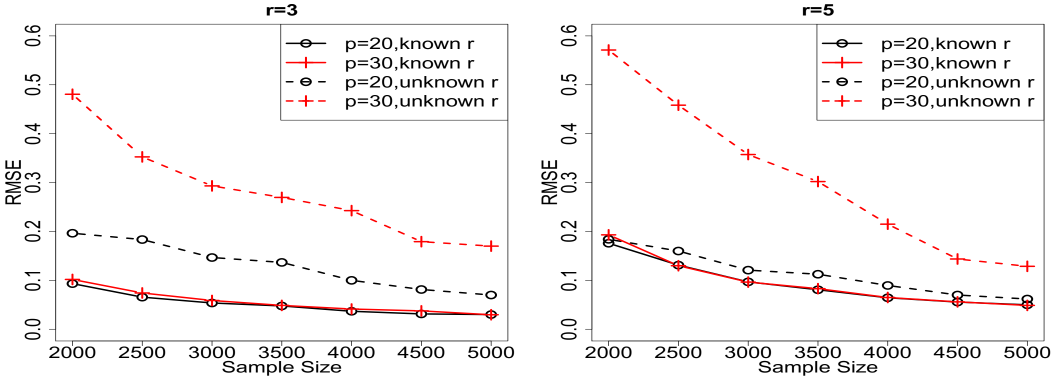
We provide simulation results on the performance of ISLET when the parameter $\boldsymbol{\mathcal{A}}$ is approximately low rank. Specifically, we first simulate the exact low Tucker rank tensor $\boldsymbol{\mathcal{A}}_0$ in the same way as the one in previous settings and simulate $\boldsymbol{\mathcal{Z}}$ as the perturbation tensor with i.i.d. standard normal entries. Then we set $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}} + \frac{\tau \|\boldsymbol{\mathcal{A}}\|_F \boldsymbol{\mathcal{Z}}}{p^3}$. The response y_j and covariate $\boldsymbol{\mathcal{X}}_j$ are generated the same to previous settings. Let $\sigma = 5, p = 20, n = [2000, 8000], s_1 = s_2 = s_3 = 12, \tau = 0, 0.1, 0.3, 0.5$. τ here characterizes how close $\boldsymbol{\mathcal{A}}$ is to the exact low-rank tensor – $\boldsymbol{\mathcal{A}}$ is exact low rank if $\tau = 0$. We apply ISLET in both the regular and sparse regimes with the tuning parameter selection scheme described in Algorithms 7 and 8 The results are collected in the Figure 14. We can see that the estimation error decreases as τ decreases or n increases; generally speaking, ISLET achieve good performance under both the regular and sparse regime when the true parameter $\boldsymbol{\mathcal{A}}$ is only approximately



(a) r = 3

(b) r = 5

Figure 12: ISLET: known rank vs unknown rank. Here, $\sigma = 5$, $\boldsymbol{r}_{ini} = \lfloor \boldsymbol{p}/3 \rfloor$.



(a) r = 3

(b) r = 5

Figure 13: Sparse ISLET: known rank vs unknown rank. Here, $\sigma = 5$; $\boldsymbol{r}_{ini} = \lfloor \boldsymbol{p}/3 \rfloor$, $s = 12$

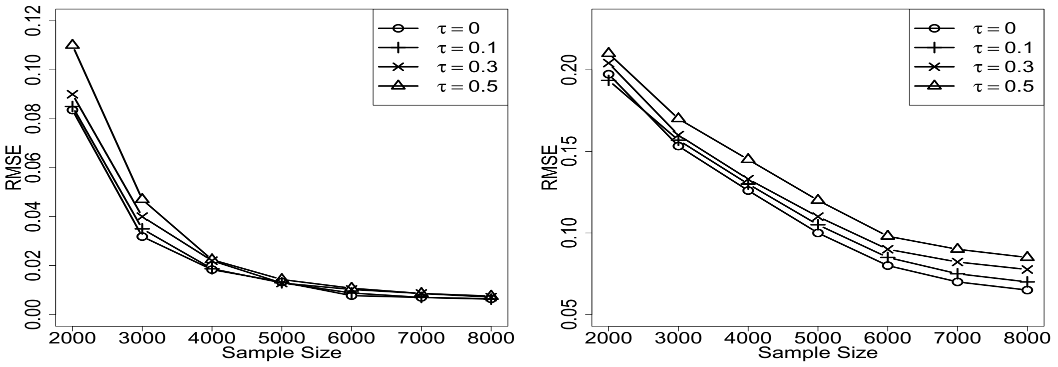
low rank.

F Proofs

We collect all proofs of the main technical results in this section.

F.1 Proof of Theorem 2

This theorem aims to develop a deterministic error bound for $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\rm HS}^2$ in terms of the sketching direction error θ, ρ, and error term $\|(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\varepsilon}\|_2^2$. Since the proof is long and technically challenging, we divide the whole argument into six steps for a better presentation. In Step 1, we introduce the notation to be used throughout the proof. In Step 2,



(a) Regular ISLET

(b) Sparse ISLET

Figure 14: Average estimation error of ISLET under approximate low Tucker rank case. Left panel: regular case; right panel: sparse case. Here, $\sigma = 5, p = 20, n = [2000, 8000], s_1 = s_2 = s_3 = 12, \tau = 0, 0.1, 0.3, 0.5$.

we transform the original high-dimensional low-rank tensor regression model to dimension-reduced one (21). We also rewrite the key quantities in the upper bound $\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\varepsilon}\|_2^2$ to $\|\widehat{\boldsymbol{\mathcal{B}}} - \widetilde{\boldsymbol{\mathcal{B}}}\|_{\rm HS}^2 + \sum_{k=1}^3 \|\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k\|_F^2$. In step 3, we introduce the factorization for $\boldsymbol{\mathcal{A}}$ and $\widehat{\boldsymbol{\mathcal{A}}}$. Based on this factorization and the property of orthogonal projection, in step 4, we decompose the loss $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\rm HS}$ into eight terms. In step 5, we bound some intermediate error terms in terms of θ and ρ using properties of the spectral norm and least singular value. In the last Step 6, we finish the proof by bounding each of the eight terms in Step 4 using the results in Step 2, 5, and Lemma 3.

Step 1 For simplicity, we denote

$$\mathbf{x}_j = \mathrm{vec}(\boldsymbol{\mathcal{X}}_j) \in \mathbb{R}^{p_1p_2p_3}, \quad \mathbf{X}_{jk} = \mathcal{M}_k(\boldsymbol{\mathcal{X}}_j) \in \mathbb{R}^{p_k \times (p_{k+1}p_{k+2})},$$

$$\mathbf{a} = \mathrm{vec}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{p_1p_2p_3}, \quad \mathbf{A}_k = \mathcal{M}_k(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{p_k \times (p_{k+1}p_{k+2})}$$

as the vectorized and matricized tensor covariates and parameter. (Note that $\mathbf{X}_{jk}$ is a matrix rather than the (j,k)-th entry of $\mathbf{X}$. Instead, we use $\mathbf{X}_{[j,k]}$ to denote the specific (i,j)-th entry of the matrix $\mathbf{X}$ in our notation system.) All mode indices $(\cdot)_k$ are in module-3, e.g., $p_4 = p_1$, $\mathbf{A}_4 = \mathbf{A}_1$, $\mathbf{X}_{j5} = \mathbf{X}_{j2}$, etc. Recall

$$\mathbf{W}_1 = (\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{V}_1, \quad \mathbf{W}_2 = (\mathbf{U}_3 \otimes \mathbf{U}_1)\mathbf{V}_2, \quad \mathbf{W}_3 = (\mathbf{U}_2 \otimes \mathbf{U}_1)\mathbf{V}_3,$$

$$\widetilde{\mathbf{W}}_1 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1, \quad \widetilde{\mathbf{W}}_2 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_2, \quad \widetilde{\mathbf{W}}_3 = (\widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_3.$$

Define

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{B}}} &= \left[\!\left[\boldsymbol{\mathcal{A}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top\right]\!\right] = \left[\!\left[\boldsymbol{\mathcal{S}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top\right]\!\right] \in \mathbb{R}^{r_1 \times r_2 \times r_3};\\
\widetilde{\mathbf{D}}_1 &= \widetilde{\mathbf{U}}_{1\perp}^\top \mathcal{M}_1(\boldsymbol{\mathcal{A}} \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3)\widetilde{\mathbf{V}}_1 \overset{\text{Lemma 1}}{=} \widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{A}_1 \widetilde{\mathbf{W}}_1 \in \mathbb{R}^{(p_1-r_1)\times r_1},\\
\widetilde{\mathbf{D}}_2 &= \widetilde{\mathbf{U}}_{2\perp}^\top \mathcal{M}_2(\boldsymbol{\mathcal{A}} \times_1 \widetilde{\mathbf{U}}_1^\top \times_3 \widetilde{\mathbf{U}}_3)\widetilde{\mathbf{V}}_2 = \widetilde{\mathbf{U}}_{2\perp}^\top \mathbf{A}_2 \widetilde{\mathbf{W}}_2 \in \mathbb{R}^{(p_2-r_2)\times r_2},\\
\widetilde{\mathbf{D}}_3 &= \widetilde{\mathbf{U}}_{3\perp}^\top \mathcal{M}_3(\boldsymbol{\mathcal{A}} \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_3 = \widetilde{\mathbf{U}}_{3\perp}^\top \mathbf{A}_3 \widetilde{\mathbf{W}}_3 \in \mathbb{R}^{(p_3-r_3)\times r_3}.
\end{aligned} \tag{38}$$

Intuitively speaking, $\widetilde{\mathbf{B}}$ is the parameter core tensor lying in the singular subspaces $\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{D}}_1, \widetilde{\mathbf{D}}_2, \widetilde{\mathbf{D}}_3$ are the parameter matrices corresponding to the arm-minus-body part lying in the singular subspace of $\mathcal{R}_1\left(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp}\right)$, $\mathcal{R}_2\left(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp}\right)$, $\mathcal{R}_3\left(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp}\right)$.

Step 2 In this step, we introduce an important decomposition for y_j and the error term $\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{\varepsilon}}\|_2^2$. In correspondence to $\widehat{\boldsymbol{\gamma}}$ (7), we construct $\widetilde{\boldsymbol{\gamma}}$ as

$$\widetilde{\boldsymbol{\gamma}} = \left(\operatorname{vec}(\widetilde{\boldsymbol{\mathcal{B}}})^\top, \operatorname{vec}(\widetilde{\mathbf{D}}_1)^\top, \operatorname{vec}(\widetilde{\mathbf{D}}_2)^\top, \operatorname{vec}(\widetilde{\mathbf{D}}_3)^\top\right)^\top \in \mathbb{R}^m. \tag{39}$$

Then for $j = 1, \ldots, n$, the response y_j can be decomposed as

$$\begin{aligned}
y_j &= \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}} \rangle + \varepsilon_j = \langle \mathbf{x}_j, \mathbf{a} \rangle + \varepsilon_j\\
&= \left\langle \mathbf{x}_j, P_{\widetilde{\mathbf{U}}}\mathbf{a} \right\rangle + \varepsilon_j + \left\langle \mathbf{x}_j, P_{\widetilde{\mathbf{U}}_\perp}\mathbf{a} \right\rangle\\
&= \left\langle \mathbf{x}_j, P_{\widetilde{\mathbf{U}}_1 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_3}\mathbf{a} \right\rangle + \sum_{k=1}^{3} \left\langle \mathbf{x}_j, P_{\mathcal{R}_k\left(\widetilde{\mathbf{U}}_{k\perp} \otimes \widetilde{\mathbf{W}}_k\right)}\mathbf{a} \right\rangle + \widetilde{\varepsilon}_j\\
&\overset{(38)}{=} \left\langle (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)^\top \mathbf{x}_j,\ (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)^\top \mathbf{a} \right\rangle\\
&\quad + \sum_{k=1}^{3} \left\langle \widetilde{\mathbf{U}}_{k\perp}^\top \mathbf{X}_{jk} \widetilde{\mathbf{W}}_k, \widetilde{\mathbf{U}}_{k\perp}^\top \mathbf{A}_k \widetilde{\mathbf{W}}_k \right\rangle + \widetilde{\varepsilon}_j\\
&\overset{(38)}{=} (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[j,:]}\operatorname{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) + \sum_{k=1}^{3} (\widetilde{\mathbf{X}}_{\mathbf{D}_k})_{[j,:]}\operatorname{vec}(\widetilde{\mathbf{D}}_k) + \widetilde{\varepsilon}_j = \widetilde{\mathbf{X}}_{[j,:]} \cdot \widetilde{\boldsymbol{\gamma}} + \widetilde{\varepsilon}_j.
\end{aligned} \tag{40}$$

Given the definitions of $\widehat{\mathbf{D}}_k$, $\widehat{\boldsymbol{\mathcal{B}}}$ (38) and $\widehat{\boldsymbol{\gamma}}$ (7) and the fact that $\widetilde{\mathbf{X}}$ is non-singular, $\widehat{\boldsymbol{\gamma}}$ can be rewritten into the following vectorized form,

$$\begin{aligned}
\widehat{\boldsymbol{\gamma}} &= \underset{\boldsymbol{\gamma} \in \mathbb{R}^m}{\arg\min} \sum_{i=1}^{n} \left(y_i - \widetilde{\mathbf{X}}_{[i,:]}\boldsymbol{\gamma}\right)^2 = \underset{\boldsymbol{\gamma} \in \mathbb{R}^m}{\arg\min} \left\| y - \widetilde{\mathbf{X}}\boldsymbol{\gamma} \right\|_2^2\\
&= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top y = \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \left(\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\gamma}} + \widetilde{\boldsymbol{\varepsilon}}\right)\\
&= \widetilde{\boldsymbol{\gamma}} + \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\boldsymbol{\varepsilon}}.
\end{aligned}$$

where $m = r_1 r_2 r_3 + \sum_{k=1}^{3}(p_k - r_k)r_k$. Thus, by the definition of $\widetilde{\gamma}$ (39), $\widehat{\gamma}$ (7), $\widehat{\boldsymbol{\mathcal{B}}}$ and $\widehat{\mathbf{D}}_k$ (8), we have

$$\|\widehat{\boldsymbol{\mathcal{B}}} - \widetilde{\boldsymbol{\mathcal{B}}}\|_{\mathrm{HS}}^2 + \sum_{k=1}^{3} \|\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k\|_F^2 = \|\widehat{\gamma} - \widetilde{\gamma}\|_2^2 = \left\|(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\varepsilon}\right\|_2^2 := \kappa^2. \tag{41}$$

Step 3 In this step, we introduce the factorization for $\boldsymbol{\mathcal{A}}$ (43). Since the left and right singular subspaces of $\mathbf{A}_k$ are $\mathbf{U}_k$ and $\mathbf{W}_k$, respectively,

$$\begin{aligned}
&\sigma_{r_k}\left(\widetilde{\mathbf{U}}_k^\top \mathbf{A}_k \widetilde{\mathbf{W}}_k\right) = \sigma_{r_k}\left(\widetilde{\mathbf{U}}_k^\top P_{\mathbf{U}_k} \mathbf{A}_k P_{\mathbf{W}_k} \widetilde{\mathbf{W}}_k\right) = \sigma_{r_k}\left((\widetilde{\mathbf{U}}_k^\top \mathbf{U}_k)\mathbf{U}_k^\top \mathbf{A}_k \mathbf{W}_k (\mathbf{W}_k^\top \widetilde{\mathbf{W}}_k)\right)\\
\geq& \sigma_{\min}(\widetilde{\mathbf{U}}_k^\top \mathbf{U}_k) \cdot \sigma_{\min}(\mathbf{U}_k^\top \mathbf{A}_k \mathbf{W}_k) \cdot \sigma_{\min}(\mathbf{W}_k^\top \widetilde{\mathbf{W}}_k)\\
=& \sqrt{1 - \|\sin\Theta(\widetilde{\mathbf{U}}_k, \mathbf{U}_k)\|^2} \cdot \sigma_{r_k}(\mathbf{A}_k) \cdot \sqrt{1 - \|\sin\Theta(\widetilde{\mathbf{W}}_k, \mathbf{W}_k)\|^2}\\
\geq& \sigma_{r_k}(\mathbf{A}_k)(1 - \theta^2) > 0.
\end{aligned} \tag{42}$$

Here, the last but one equality is due to the property of $\sin\Theta$ distance (c.f., Lemma 1 in [19]). Thus, $\mathrm{rank}(\widetilde{\mathbf{U}}_k^\top \mathbf{A}_k \widetilde{\mathbf{W}}_k) = r_k$, which is a full rank matrix. Thus,

$$\begin{aligned}
&\boldsymbol{\mathcal{A}} = \llbracket \boldsymbol{\mathcal{B}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \rrbracket\\
=& \left\llbracket \llbracket \boldsymbol{\mathcal{B}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \rrbracket; \mathbf{U}_1(\widetilde{\mathbf{U}}_1^\top \mathbf{U}_1)^{-1}\widetilde{\mathbf{U}}_1^\top, \mathbf{U}_2(\widetilde{\mathbf{U}}_2^\top \mathbf{U}_2)^{-1}\widetilde{\mathbf{U}}_2^\top, \mathbf{U}_3(\widetilde{\mathbf{U}}_3^\top \mathbf{U}_3)^{-1}\widetilde{\mathbf{U}}_3^\top \right\rrbracket\\
=& \left\llbracket \boldsymbol{\mathcal{A}}; \mathbf{U}_1(\widetilde{\mathbf{U}}_1^\top \mathbf{U}_1)^{-1}\widetilde{\mathbf{U}}_1^\top, \mathbf{U}_2(\widetilde{\mathbf{U}}_2^\top \mathbf{U}_2)^{-1}\widetilde{\mathbf{U}}_2^\top, \mathbf{U}_3(\widetilde{\mathbf{U}}_3^\top \mathbf{U}_3)^{-1}\widetilde{\mathbf{U}}_3^\top \right\rrbracket\\
=& \left\llbracket \boldsymbol{\mathcal{A}}; \mathbf{A}_1\widetilde{\mathbf{W}}_1(\widetilde{\mathbf{U}}_1^\top \mathbf{A}_1 \widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{U}}_1^\top, \mathbf{A}_2\widetilde{\mathbf{W}}_2(\widetilde{\mathbf{U}}_2^\top \mathbf{A}_2 \widetilde{\mathbf{W}}_2)^{-1}\widetilde{\mathbf{U}}_2^\top, \mathbf{A}_3\widetilde{\mathbf{W}}_3(\widetilde{\mathbf{U}}_3^\top \mathbf{A}_3 \widetilde{\mathbf{W}}_3)^{-1}\widetilde{\mathbf{U}}_3^\top \right\rrbracket
\end{aligned} \tag{43}$$

The fourth equality is because the left singular space and right singular space of $\mathbf{A}_k$ is $\mathbf{U}_k$ and $\mathbf{W}_k$.

Recall

$$\widehat{\boldsymbol{\mathcal{A}}} = \left\llbracket \widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{L}}_1, \widehat{\mathbf{L}}_2, \widehat{\mathbf{L}}_3 \right\rrbracket, \quad \widehat{\mathbf{L}}_k = (\widetilde{\mathbf{U}}_k \widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k + \widetilde{\mathbf{U}}_{k\perp}\widehat{\mathbf{D}}_k)(\widehat{\mathbf{B}}_k \widetilde{\mathbf{V}}_k)^{-1}, \quad k = 1, 2, 3.$$

Denote $\widetilde{\mathbf{B}}_k = \mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{B}}})$, $\widehat{\mathbf{B}}_k = \mathcal{M}_k(\widehat{\boldsymbol{\mathcal{B}}})$. In parallel to the definition of $\widehat{\mathbf{L}}_k$, we define

$$\begin{aligned}
\widetilde{\mathbf{L}}_1 =& (\widetilde{\mathbf{U}}_1 \widetilde{\mathbf{B}}_1 \widetilde{\mathbf{V}}_1 + \widetilde{\mathbf{U}}_{1\perp}\widetilde{\mathbf{D}}_1)(\widetilde{\mathbf{B}}_1 \widetilde{\mathbf{V}}_1)^{-1},\\
=& \left(\widetilde{\mathbf{U}}_1 \widetilde{\mathbf{U}}_1^\top \mathbf{A}_1 (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1 + \widetilde{\mathbf{U}}_{1\perp}\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{A}_1 (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1\right)\\
&\cdot \left(\widetilde{\mathbf{U}}_1^\top \mathbf{A}_1 (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1\right)^{-1}\\
=& \mathbf{A}_1 \widetilde{\mathbf{W}}_1 \left(\widetilde{\mathbf{U}}_1^\top \mathbf{A}_1 \widetilde{\mathbf{W}}_1\right)^{-1}.
\end{aligned} \tag{44}$$

Similarly,

$$\begin{aligned}\widetilde{\mathbf{L}}_2 =&(\widetilde{\mathbf{U}}_2\widetilde{\mathbf{B}}_2\widetilde{\mathbf{V}}_2 + \widetilde{\mathbf{U}}_{2\perp}\widetilde{\mathbf{D}}_2)(\widetilde{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1} = \mathbf{A}_2\widetilde{\mathbf{W}}_2\left(\widetilde{\mathbf{U}}_2^\top\mathbf{A}_2\widetilde{\mathbf{W}}_2\right)^{-1},\\ \widetilde{\mathbf{L}}_3 =&(\widetilde{\mathbf{U}}_3\widetilde{\mathbf{B}}_3\widetilde{\mathbf{V}}_3 + \widetilde{\mathbf{U}}_{3\perp}\widetilde{\mathbf{D}}_3)(\widetilde{\mathbf{B}}_3\widetilde{\mathbf{V}}_3)^{-1} = \mathbf{A}_3\widetilde{\mathbf{W}}_3\left(\widetilde{\mathbf{U}}_3^\top\mathbf{A}_3\widetilde{\mathbf{W}}_3\right)^{-1}.\end{aligned}$$

Thus, in addition to $\widehat{\boldsymbol{\mathcal{A}}} = [\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{L}}_1, \widehat{\mathbf{L}}_2, \widehat{\mathbf{L}}_3]\!]$, we have

$$\boldsymbol{\mathcal{A}} = [\![\widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{L}}_1, \widetilde{\mathbf{L}}_2, \widetilde{\mathbf{L}}_3]\!] \tag{45}$$

Step 4 Next, we analyze the estimation error of $\widehat{\boldsymbol{\mathcal{A}}}$. First, the error bound of $\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}$ can be decomposed into eight parts,

$$\begin{aligned}&\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}^2 = \left\|[\![\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1} + P_{\widetilde{\mathbf{U}}_{1\perp}}, P_{\widetilde{\mathbf{U}}_2} + P_{\widetilde{\mathbf{U}}_{2\perp}}, P_{\widetilde{\mathbf{U}}_3} + P_{\widetilde{\mathbf{U}}_{3\perp}}]\!]\right\|_{\mathrm{HS}}^2\\ =&\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2 + \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2\\ &+ \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_{2\perp}^\top, \widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2 + \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_{\mathrm{HS}}^2\\ &+ \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_{2\perp}^\top, \widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_{\mathrm{HS}}^2 + \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_{\mathrm{HS}}^2\\ &+ \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^\top, \widetilde{\mathbf{U}}_{2\perp}^\top, \widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2 + \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^\top, \widetilde{\mathbf{U}}_{2\perp}^\top, \widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_{\mathrm{HS}}^2.\end{aligned} \tag{46}$$

Here we used the fact that $P_{\widetilde{\mathbf{U}}_1}$ and $P_{\widetilde{\mathbf{U}}_{1\perp}}$ are orthogonal complementary. We aim to apply Lemma 3 to analyze each term above in the next two steps.

Step 5 Before giving the upper bounds for each term of (46), we denote

$$\begin{aligned}\lambda_k =&\max\left\{\left\|\widehat{\mathbf{D}}_k(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\|, \left\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\|\right\},\\ \pi_k =&\|(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\widetilde{\mathbf{B}}_k\|, \quad k = 1, 2, 3\end{aligned} \tag{47}$$

and aim to provide upper bounds for λ_k, π_k in this step. By definition of $\widetilde{\mathbf{B}}_k$ and the fact that the right singular vector of $\mathbf{A}_k$ is $\mathbf{W}_k$,

$$\begin{aligned}\pi_1 =&\left\|(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\widetilde{\mathbf{B}}_1\right\| = \left\|(\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\right\|\\ \leq&\left\|\left(\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1\widetilde{\mathbf{W}}_1\right)^{-1}\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1\right\| = \left\|\left(\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1\mathbf{W}_1\mathbf{W}_1^\top\widetilde{\mathbf{W}}_1\right)^{-1}\widetilde{\mathbf{U}}_1^\top\mathbf{A}_1\mathbf{W}_1\right\|\\ \leq&\left\|(\mathbf{W}_1^\top\widetilde{\mathbf{W}}_1)^{-1}\right\| = \sigma_{\min}^{-1}(\widetilde{\mathbf{W}}_1^\top\mathbf{W}_1) = \left(1 - \|\sin\Theta(\widetilde{\mathbf{W}}_k, \mathbf{W}_k)\|^2\right)^{-1/2}\\ \leq&\frac{1}{(1-\theta^2)^{1/2}}.\end{aligned} \tag{48}$$

Similarly, the same upper bounds also applies to π_2 and π_3.

Based on definitions of $\widetilde{\mathbf{D}}_k$ and $\widetilde{\mathbf{B}}_k$ and the fact that the left singular subspace of $\mathbf{A}_k$ is $\mathbf{U}_k$, we have

$$\begin{aligned}
&\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\|^2+1 = \left\|\widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k)^{-1}\right\|^2+1\\
=&\left\|\begin{bmatrix}\mathbf{I}_{r_k}\\ \widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k)^{-1}\end{bmatrix}\right\|^2 = \left\|\begin{bmatrix}\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k)^{-1}\\ \widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k)^{-1}\end{bmatrix}\right\|^2\\
=&\left\|\mathbf{A}_k\widetilde{\mathbf{W}}_k\left(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k\right)^{-1}\right\|^2 = \left\|\mathbf{U}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k\left(\widetilde{\mathbf{U}}_k^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k\right)^{-1}\right\|^2 \qquad (49)\\
=&\left\|\mathbf{U}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k\left(\mathbf{U}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k\right)^{-1}\left(\widetilde{\mathbf{U}}_k^\top\mathbf{U}_k\right)^{-1}\right\|^2\\
=&\left\|\left(\widetilde{\mathbf{U}}_1^\top\mathbf{U}_1\right)^{-1}\right\|^2 = \sigma_{\min}^{-2}\left(\widetilde{\mathbf{U}}_1^\top\mathbf{U}_1\right) = \left(1-\|\sin\Theta(\widetilde{\mathbf{U}}_1,\mathbf{U}_1)\|^2\right)^{-1}\leq\frac{1}{1-\theta^2},
\end{aligned}$$

which implies

$$\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\|\leq\sqrt{\frac{1}{1-\theta^2}-1}=\sqrt{\frac{\theta^2}{1-\theta^2}}.$$

By the assumption of the theorem that $\|\widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\|\leq\rho$ and $\theta\leq 1/2$, we have

$$\lambda_k\leq\max\left\{\rho,\frac{\theta}{\sqrt{1-\theta^2}}\right\}\leq\rho+\frac{2}{\sqrt{3}}\theta,\quad k=1,2,3. \qquad (50)$$

Step 6 Now we are ready to give upper bounds for all terms in (46).

- First, by definition of $\widehat{\boldsymbol{\mathcal{B}}}$, $\widehat{\boldsymbol{\mathcal{A}}}$ (9),

$$\begin{aligned}
&[\![\widehat{\boldsymbol{\mathcal{A}}};\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top]\!] = \left[\!\!\left[[\![\widehat{\boldsymbol{\mathcal{B}}};\widehat{\mathbf{L}}_1,\widehat{\mathbf{L}}_2,\widehat{\mathbf{L}}_3]\!];\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top\right]\!\!\right]\\
=&\left[\!\!\left[\widehat{\boldsymbol{\mathcal{B}}};\widetilde{\mathbf{U}}_1^\top\widehat{\mathbf{L}}_1,\widetilde{\mathbf{U}}_2^\top\widehat{\mathbf{L}}_2,\widetilde{\mathbf{U}}_3^\top\widehat{\mathbf{L}}_3\right]\!\!\right].
\end{aligned} \qquad (51)$$

Here,

$$\widetilde{\mathbf{U}}_k^\top\widehat{\mathbf{L}}_k=\widetilde{\mathbf{U}}_k^\top\left((\widetilde{\mathbf{U}}_k\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k+\widetilde{\mathbf{U}}_{k\perp}\widehat{\mathbf{D}}_k)(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right)=(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}=\mathbf{I}_{r_k}.$$

Similarly, we have $\widetilde{\mathbf{U}}_k^\top\widetilde{\mathbf{L}}_k=\mathbf{I}_{r_k}$.
Thus, $[\![\widehat{\boldsymbol{\mathcal{A}}};\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top]\!]=\widehat{\boldsymbol{\mathcal{B}}}$. By definition of $\widetilde{\boldsymbol{\mathcal{B}}}$ (38), we have

$$\begin{aligned}
\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}});\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2=&\left\|[\![\widehat{\boldsymbol{\mathcal{A}}};\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top]\!]-[\![\boldsymbol{\mathcal{A}};\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top]\!]\right\|_{\mathrm{HS}}^2\\
=&\|\widehat{\boldsymbol{\mathcal{B}}}-\widetilde{\boldsymbol{\mathcal{B}}}\|_{\mathrm{HS}}^2.
\end{aligned} \qquad (52)$$

- Note that

$$\begin{aligned}
& \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^{\top}, \widetilde{\mathbf{U}}_2^{\top}, \widetilde{\mathbf{U}}_3^{\top}]\!]\right\|_{\mathrm{HS}}^2 \\
\overset{(45),(51)}{=} & \left\|[\![\widehat{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{U}}_{1\perp}^{\top}\widehat{\mathbf{L}}_1, \widetilde{\mathbf{U}}_2^{\top}\widehat{\mathbf{L}}_2, \widetilde{\mathbf{U}}_3^{\top}\widehat{\mathbf{L}}_3]\!] - [\![\widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{U}}_{1\perp}^{\top}\widetilde{\mathbf{L}}_1, \widetilde{\mathbf{U}}_2^{\top}\widetilde{\mathbf{L}}_2, \widetilde{\mathbf{U}}_3^{\top}\widetilde{\mathbf{L}}_3]\!]\right\|_{\mathrm{HS}}^2 \\
\overset{(9)(44)}{=} & \left\|[\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \mathbf{I}, \mathbf{I}]\!] - [\![\widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \mathbf{I}, \mathbf{I}]\!]\right\|_{\mathrm{HS}}^2 \\
\overset{\text{Lemma 1}}{=} & \left\|\widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\widehat{\mathbf{B}}_1 - \widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\widetilde{\mathbf{B}}_1\right\|_F^2
\end{aligned} \tag{53}$$

  By the first part of Lemma 3,

$$\begin{aligned}
& \left\|\widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\widehat{\mathbf{B}}_1 - \widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}\widetilde{\mathbf{B}}_1\right\|_F^2 \\
\leq & \left(\pi_1\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F + \lambda_1\|\widehat{\mathbf{B}}_1 - \widetilde{\mathbf{B}}_1\|_F + \pi_1\lambda_1\|\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1 - \widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1\|_F\right)^2 \\
\overset{(48)(50)}{\leq} & \left(\frac{1}{\sqrt{1-\theta^2}}\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F + (\rho + \frac{2}{\sqrt{3}}\theta)\kappa + (\rho + \frac{2}{\sqrt{3}}\theta)\frac{1}{\sqrt{1-\theta^2}}\kappa\right)^2 \\
\leq & \frac{1}{1-\theta^2}\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F^2 + C_1(\rho+\theta)\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F\kappa + C_2(\rho+\theta)^2\kappa^2 \\
\leq & \|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F^2 + 2\theta^2\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F^2 + C_1(\rho+\theta)\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F\kappa + C_2(\rho+\theta)^2\kappa^2 \\
\leq & \|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F^2 + C(\rho+\theta)\kappa^2.
\end{aligned}$$

  Here, the last inequality is due to the fact that $\|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F \leq \kappa$. Therefore,

$$\begin{aligned}
& \left\|\left[\!\!\left[(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^{\top}, \widetilde{\mathbf{U}}_2^{\top}, \widetilde{\mathbf{U}}_3\right]\!\!\right]\right\|_{\mathrm{HS}}^2 \leq \|\widehat{\mathbf{D}}_1 - \widetilde{\mathbf{D}}_1\|_F^2 + C(\rho+\theta)\kappa^2; \\
\text{similarly} \quad & \left\|\left[\!\!\left[(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^{\top}, \widetilde{\mathbf{U}}_{2\perp}^{\top}, \widetilde{\mathbf{U}}_3\right]\!\!\right]\right\|_{\mathrm{HS}}^2 \leq \|\widehat{\mathbf{D}}_2 - \widetilde{\mathbf{D}}_2\|_F^2 + C(\rho+\theta)\kappa^2, \\
& \left\|\left[\!\!\left[(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_1^{\top}, \widetilde{\mathbf{U}}_2^{\top}, \widetilde{\mathbf{U}}_{3\perp}\right]\!\!\right]\right\|_{\mathrm{HS}}^2 \leq \|\widehat{\mathbf{D}}_3 - \widetilde{\mathbf{D}}_3\|_F^2 + C(\rho+\theta)\kappa^2.
\end{aligned} \tag{54}$$

- By similar argument as (53), we have

$$\begin{aligned}
& \left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}); \widetilde{\mathbf{U}}_{1\perp}^{\top}, \widetilde{\mathbf{U}}_{2\perp}^{\top}, \widetilde{\mathbf{U}}_3]\!]\right\|_F^2 \\
= & \left\|[\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \widehat{\mathbf{D}}_2(\widehat{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1}, \mathbf{I}]\!] - [\![\widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \widetilde{\mathbf{D}}_2(\widetilde{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1}, \mathbf{I}]\!]\right\|_F^2
\end{aligned}$$

  By the second part of Lemma 3,

$$\begin{aligned}
& \left\|[\![\widehat{\boldsymbol{\mathcal{B}}}; \widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \widehat{\mathbf{D}}_2(\widehat{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1}, \mathbf{I}]\!] - [\![\widetilde{\boldsymbol{\mathcal{B}}}; \widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1}, \widetilde{\mathbf{D}}_2(\widetilde{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1}, \mathbf{I}]\!]\right\|_F^2 \\
\leq & \left(\lambda_1\lambda_2\|\widehat{\boldsymbol{\mathcal{B}}} - \widetilde{\boldsymbol{\mathcal{B}}}\|_F + \sum_{k=1,2}\pi_k\lambda_1\lambda_2/\lambda_k\|\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k\|_F + \sum_{k=1,2}\pi_k\lambda_1\lambda_2\|\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k - \widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k\|_F\right)^2 \\
\overset{(41)}{\leq} & (\lambda_1\lambda_2 + \pi_1\lambda_2 + \pi_2\lambda_1 + \pi_1\lambda_1\lambda_2 + \pi_2\lambda_1\lambda_2)^2\kappa^2 \overset{(48)}{\leq} C(\rho+\theta)^2\kappa^2.
\end{aligned}$$

Therefore,

$$\begin{aligned} &\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}});\widetilde{\mathbf{U}}_{1\perp}^\top,\widetilde{\mathbf{U}}_{2\perp}^\top,\widetilde{\mathbf{U}}_{3}^\top]\!]\right\|_F^2 \leq C(\rho+\theta)^2\kappa^2;\\ \text{similarly,}\quad &\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}});\widetilde{\mathbf{U}}_{1\perp}^\top,\widetilde{\mathbf{U}}_{2}^\top,\widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_F^2 \leq C(\rho+\theta)^2\kappa^2, \\ &\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}});\widetilde{\mathbf{U}}_{1}^\top,\widetilde{\mathbf{U}}_{2\perp}^\top,\widetilde{\mathbf{U}}_{3\perp}^\top]\!]\right\|_F^2 \leq C(\rho+\theta)^2\kappa^2. \end{aligned} \tag{55}$$

- By the second part of Lemma 3,

$$\begin{aligned} &\left\|[\![(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}});\widetilde{\mathbf{U}}_{1\perp}^\top,\widetilde{\mathbf{U}}_{2\perp}^\top,\widetilde{\mathbf{U}}_{3\perp}]\!]\right\|_F^2\\ =&\Big\|[\![\widehat{\boldsymbol{\mathcal{B}}};\widehat{\mathbf{D}}_1(\widehat{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1},\widehat{\mathbf{D}}_2(\widehat{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1},\widehat{\mathbf{D}}_3(\widehat{\mathbf{B}}_3\widetilde{\mathbf{V}}_3)^{-1}]\!]\\ &-[\![\widetilde{\boldsymbol{\mathcal{B}}};\widetilde{\mathbf{D}}_1(\widetilde{\mathbf{B}}_1\widetilde{\mathbf{V}}_1)^{-1},\widetilde{\mathbf{D}}_2(\widetilde{\mathbf{B}}_2\widetilde{\mathbf{V}}_2)^{-1},\widetilde{\mathbf{D}}_3(\widetilde{\mathbf{B}}_3\widetilde{\mathbf{V}}_3)^{-1}]\!]\Big\|_F^2\\ \leq&\Big(\lambda_1\lambda_2\lambda_3\|\widehat{\boldsymbol{\mathcal{B}}}-\widetilde{\boldsymbol{\mathcal{B}}}\|_F+\sum_{k=1,2,3}\pi_k\lambda_1\lambda_2\lambda_3/\lambda_k\|\widehat{\mathbf{D}}_k-\widetilde{\mathbf{D}}_k\|_F\\ &+\sum_{k=1,2,3}\pi_k\lambda_1\lambda_2\lambda_3\|\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k-\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k\|_F\Big)^2\\ \overset{(41)(48)}{\leq}&C(\rho+\theta)^4\kappa^2. \end{aligned} \tag{56}$$

Combining (46), (52), (54), (55) and (56), we finally have

$$\left\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \leq \|\widehat{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}\|_F^2+\sum_{k=1}\|\widehat{\mathbf{D}}_k-\widetilde{\mathbf{D}}_k\|_F^2+C(\rho+\theta)\kappa^2=(1+C(\rho+\theta))\kappa^2.$$

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

F.2 Proof of Theorem 3

This theorem gives a deterministic error bound of $\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}^2$ in terms of θ,ρ and $\|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\mathbf{X}_{\boldsymbol{\mathcal{B}}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{B}}\|_2^2$, $\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top\widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1}\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2$, $\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k,[:,G_i^k]})^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n\|_2^2$ for the sparse ISLET estimator $\widehat{\boldsymbol{\mathcal{A}}}$ in the sparse low-rank tensor regression model. To prove this theorem, we first rewrite the original high-dimensional regression model to four dimension-reduced ones (59), (60). Then we derive error bounds for the least square estimator or group Lasso estimator in terms of $\|\widehat{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}\|_{\mathrm{HS}}^2$ or $\|\widehat{\mathbf{E}}_k-\widetilde{\mathbf{E}}_k\|_F^2$ for each of these dimension-reduced regression models. The rest of the proof aims to assemble the upper bound for $\|\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}^2$, which essentially follows from Steps 3-6 in the proof of Theorem 2.

Denote

$$\mathbf{A}_k=\mathcal{M}_k(\boldsymbol{\mathcal{A}}),\quad \mathbf{a}=\mathrm{vec}(\boldsymbol{\mathcal{A}}),\quad \mathbf{X}_{jk}=\mathcal{M}_k(\boldsymbol{\mathcal{X}}_j),\quad \mathbf{x}_j=\mathrm{vec}(\boldsymbol{\mathcal{X}}_j),\quad 1\leq j\leq n,\quad k=1,2,3;$$

$$\begin{aligned}&\widetilde{\boldsymbol{\mathcal{B}}} = [\![\boldsymbol{\mathcal{A}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!];\\ &\widetilde{\mathbf{E}}_k = \mathcal{M}_k(\boldsymbol{\mathcal{A}} \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top)\widetilde{\mathbf{V}}_k = \mathbf{A}_k\widetilde{\mathbf{W}}_k \in \mathbb{R}^{p_k\times r_k}, \quad k=1,2,3;\end{aligned} \tag{57}$$

$$\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} = \mathrm{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) \in \mathbb{R}^{p_1p_2p_3}, \quad \widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k} = \mathrm{vec}(\widetilde{\mathbf{E}}_k) \in \mathbb{R}^{p_kr_k}, \quad k=1,2,3. \tag{58}$$

Then similarly as the argument (40) in the proof of Theorem 2, we can write down the following partial regression formulas that relate y_j and $(\boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}})$,

$$\begin{aligned} y_j =& \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}}\rangle + \varepsilon_j = \langle \mathbf{x}_j, \mathbf{a}\rangle + \varepsilon_j\\ =& \left\langle \mathbf{x}_j, P_{\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1}\mathbf{a}\right\rangle + \varepsilon_j + \langle \mathbf{x}_j, P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathbf{a}\rangle\\ =& \left\langle (\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)^\top\mathbf{x}_j, (\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)^\top\mathbf{a}\right\rangle + (\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}})_j\\ \overset{(57)(58)}{=}& (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{[j,:]}\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} + (\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}})_j, \end{aligned} \tag{59}$$

$$\begin{aligned} y_j =& \langle \boldsymbol{\mathcal{X}}_j, \boldsymbol{\mathcal{A}}\rangle + \varepsilon_j\\ =& \left\langle \boldsymbol{\mathcal{X}}_j, P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})}[\boldsymbol{\mathcal{A}}]\right\rangle + \varepsilon_j + \left\langle \boldsymbol{\mathcal{X}}_j, P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}[\boldsymbol{\mathcal{A}}]\right\rangle\\ =& \left\langle \mathbf{X}_{jk}\widetilde{\mathbf{W}}_k,\ \mathbf{A}_k\widetilde{\mathbf{W}}_k\right\rangle + (\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k})_j\\ \overset{(57)(58)}{=}& (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[j,:]}\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k} + (\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k})_j \end{aligned} \tag{60}$$

for $j=1,\ldots,n$ and $k=1,2,3$. We discuss the estimation errors of $\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k}$ $(k\in J_s)$, $\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k}$ $(k\notin J_s)$, and $\widehat{\boldsymbol{\mathcal{B}}}$ separately as below.

- For any $k\in J_s$, due to the definition that

  $$\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k} = \mathrm{vec}(\widetilde{\mathbf{E}}_k), \quad \widetilde{\mathbf{E}}_k = \mathbf{A}_k\widetilde{\mathbf{W}}_k,$$

  and the left singular vectors of $\mathbf{A}_k$ is $\mathbf{U}_k$ that satisfying $\|\mathbf{U}_k\|_0 = \sum_{i=1}^{p_k} 1_{\{(\mathbf{U}_k)_{[i,:]}\neq 0\}} \leq s_k$, $\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k}$ is correspondingly group-wise sparse. More specifically, let $G_k^i = \{i, i+p_k, \ldots, i+p_k(r_k-1)\}$ with $i=1,\ldots,p_k$ be a partition of $\{1,\ldots,p_kr_k\}$. Then

  $$\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k}^i := (\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k})_{G_k^i} \in \mathbb{R}^{r_k}, \quad \sum_{i=1}^{p_k} 1_{\{\widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k}^i\neq 0\}} \leq s_k. \tag{61}$$

  Accordingly, $\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n_2\times(p_kr_k)}$ are with grouped covariates with respect to $\{G_k^1,\ldots,G_k^{p_k}\}$:

  $$\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i = (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{[:,G_k^i]} \in \mathbb{R}^{n\times r_k}, \quad i=1,\ldots,p_k. \tag{62}$$

  Recall $\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k}$ is the group Lasso estimator,

  $$\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k} = \mathop{\arg\min}_{\boldsymbol{\gamma}\in\mathbb{R}^{(p_kr_k)}} \|y - \widetilde{\mathbf{X}}_{\mathbf{E}_k}\boldsymbol{\gamma}\|_2^2 + \eta_k\sum_{i=1}^{p_k}\|\boldsymbol{\gamma}_{G_k^i}\|_2.$$

By the group-wise sparsity structure (61)(62), the partial linear regression model (60), the assumption that $\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n_2 \times (p_k r_k)}$ satisfies GRIP assumption with $\delta < 1/4$, and $\eta_k = C \max_{1\leq i \leq p_k} \|(\widetilde{\mathbf{X}}^i_{\mathbf{E}_k})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2$ for constant $C \geq 3$, Lemma 11 yields

$$\|\widehat{\mathbf{E}}_k - \widetilde{\mathbf{E}}_k\|_F = \|\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k} - \widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k}\|_2 \leq \frac{C\sqrt{s_k}\eta_k}{n} \leq C\sqrt{s_k} \max_{1\leq i\leq p_k} \|(\widetilde{\mathbf{X}}^i_{\mathbf{E}_k})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n\|_2, \quad \forall k \in J_s. \tag{63}$$

- For $k \notin J_s$, recall $\widehat{\mathbf{E}}_k$ is evaluated via the least square estimator,

$$\mathrm{vec}(\widehat{\mathbf{E}}_k) = \widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k}, \quad \widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k} = \mathop{\arg\min}_{\boldsymbol{\gamma} \in \mathbb{R}^{(p_k r_k)}} \left\| y - \widetilde{\mathbf{X}}_{\mathbf{E}_k} \boldsymbol{\gamma} \right\|_2^2.$$

By linear regression model (60) and the definition of the least square estimator,

$$\|\widehat{\mathbf{E}}_k - \widetilde{\mathbf{E}}_k\|_F = \|\widehat{\boldsymbol{\gamma}}_{\mathbf{E}_k} - \widetilde{\boldsymbol{\gamma}}_{\mathbf{E}_k}\|_2 = \left\| (\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1} \widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k} \right\|_2^2. \tag{64}$$

- In addition, recall

$$\mathrm{vec}(\widehat{\boldsymbol{\mathcal{B}}}) = \widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}}, \quad \widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} = \mathop{\arg\min}_{\boldsymbol{\gamma} \in \mathbb{R}^{r_1 r_2 r_3}} \|y - \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \boldsymbol{\gamma}\|_2^2.$$

By linear regression model (59) and the definition of the least square estimator $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}}$,

$$\|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}\|_{\mathrm{HS}}^2 = \|\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\mathcal{B}}} - \boldsymbol{\gamma}_{\boldsymbol{\mathcal{B}}}\|_2^2 = \|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1} \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}}\|_2^2. \tag{65}$$

Given $\theta = \max\{\|\sin\Theta(\widetilde{\mathbf{U}}_k, \mathbf{U}_k)\|, \|\sin\Theta(\widetilde{\mathbf{W}}_k, \mathbf{W}_k)\|\} \leq 1/2$, similarly as the proof of Theorem 2, one can show $\widetilde{\mathbf{U}}_k^\top \widetilde{\mathbf{E}}_k$ is non-singular. Therefore,

$$\begin{aligned} \|\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}\|_{\mathrm{HS}}^2 + \sum_{k=1}^{3} \|\widehat{\mathbf{E}}_k - \widetilde{\mathbf{E}}_k\|_F^2 \leq & \left\| (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1} \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}} \right\|_2^2 + C \sum_{k \in J_s} s_k \max_{1\leq i\leq p_k} \left\| (\widetilde{\mathbf{X}}^i_{\mathbf{E}_k})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n_2 \right\|_2^2 \\ & + \sum_{k \notin J_s} \left\| (\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1} \widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k} \right\|_2^2. \end{aligned}$$

The rest of the proof directly follows from Steps 3 - 6 in Theorem 3. □

F.3 Proof of Theorem 4

The goal of Theorem 4 is to give a probabilistic error bound for regular tensor regression via ISLET. The high level idea is to first derive the error bound for importance sketching regression by a perturbation bound of the HOOI outcome (Theorem 1 in [138]), and then apply the oracle inequality in Theorem 2 to obtain the final estimation error rate. For a better presentation, we divide the long proof into six steps. First in Step 1, we bound the initialization error of $\widetilde{\mathbf{U}}_k^{(0)}$ using perturbation theory [19] and concentration inequality

(Lemmas 2 and 4). Then in Step 2, we aim to apply Theorem 1 in [138] to get an error bound for the importance sketching directions $\widetilde{\mathbf{U}}_k$. The central goal of Step 3 is to prove an error bound for θ. In Steps 4, we move on to the second batch of sample and derive error bounds for a few intermediate terms. In step 5, we evaluate key quantities ρ and $\left\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{\varepsilon}}\right\|_2^2$ in the context of Theorem 2. Finally, we plug in all quantities to Theorem 2 and finish the proof.

We begin the proof by introducing some notations. Throughout the proof, the mode indices $(\cdot)_k$ are presented in modulo 3: e.g., $\mathbf{U}_4 = \mathbf{U}_1$, $\mathbf{V}_5 = \mathbf{V}_2$. For convenience, we denote

$$\widetilde{\sigma}^2 = \|\boldsymbol{\mathcal{A}}\|_{\mathrm{HS}}^2 + \sigma^2, \quad \mathbf{A}_k = \mathcal{M}_k(\boldsymbol{\mathcal{A}}), \quad \widetilde{\mathbf{A}}_k = \mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{A}}}), \quad \mathbf{X}_{ik} = \mathcal{M}_k(\boldsymbol{\mathcal{X}}_i)$$

for $k = 1, 2, 3$. $p = \max\{p_1, p_2, p_3\}$, $r = \max\{r_1, r_2, r_3\}$. To avoid repeating similar notations consecutively, throughout the proof of this theorem we slightly abuse the notation and denote

$$\mathbf{U}_{k+2} \otimes \mathbf{U}_{k+1} = \begin{cases} \mathbf{U}_3 \otimes \mathbf{U}_2, & k = 1; \\ \mathbf{U}_3 \otimes \mathbf{U}_1, & k = 2; \\ \mathbf{U}_2 \otimes \mathbf{U}_1, & k = 3 \end{cases}$$

without ambiguity. Other related notations, e.g., $(\mathbf{U}_{k+2\perp}\mathbf{V}) \otimes \mathbf{U}_{k+1}$, are defined in a similar fashion.

The rest of the proof for Theorem 4 is divided into 6 steps.

Step 1 We first develop the error bound for $\widetilde{\mathbf{U}}_1^{(0)}$, $\widetilde{\mathbf{U}}_2^{(0)}$, and $\widetilde{\mathbf{U}}_3^{(0)}$. Particularly, we aim to show that

$$\mathbb{P}\left(\left\|\sin\Theta(\widetilde{\mathbf{U}}_k^{(0)}, \mathbf{U}_k)\right\| \leq \left(\frac{C\widetilde{\sigma}\sqrt{p_k/n_1}}{\lambda_k} + \frac{\widetilde{\sigma}^2\sqrt{p_1p_2p_3}/n_1}{\lambda_k^2}\right) \wedge 1, k = 1, 2, 3\right) \geq 1 - p^{-C}. \tag{66}$$

We only focus on $\widetilde{\mathbf{U}}_1^{(0)}$ as the conclusions for $\widetilde{\mathbf{U}}_2^{(0)}$ and $\widetilde{\mathbf{U}}_3^{(0)}$ similarly follow. Recall the baseline unbiased estimator

$$\widetilde{\boldsymbol{\mathcal{A}}} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i^{(1)}\boldsymbol{\mathcal{X}}_i^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle\boldsymbol{\mathcal{X}}_i^{(1)}, \boldsymbol{\mathcal{A}}\rangle + \varepsilon_i^{(1)}\right)\boldsymbol{\mathcal{X}}_i^{(1)} \in \mathbb{R}^{p_1\times p_2\times p_3}.$$

Since the left and right singular subspaces of $\mathbf{A}_1$ are $\mathbf{U}_1$ and $\mathbf{W}_1$, respectively, we further have $\widetilde{\mathbf{A}}_1 \in \mathbb{R}^{p_1 \times (p_2 p_3)}$ and

$$
\begin{aligned}
\widetilde{\mathbf{A}}_1 =& \mathcal{M}_1\left(\widetilde{\boldsymbol{\mathcal{A}}}\right) = \frac{1}{n_1}\sum_{i=1}^{n} y_i^{(1)}\mathbf{X}_{i1}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{X}_{i1}^{(1)}, \mathbf{A}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{X}_{i1}^{(1)} \\
=& \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{X}_{i1}^{(1)}, P_{\mathbf{U}_1}\mathbf{A}_1 P_{\mathbf{W}_1}\rangle + \varepsilon_i^{(1)}\right)\mathbf{X}_{i1}^{(1)} \\
=& \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\operatorname{tr}\left((\mathbf{X}_{i1}^{(1)})^\top \mathbf{U}_1\mathbf{U}_1^\top \mathbf{A}_1\mathbf{W}_1\mathbf{W}_1^\top\right) + \varepsilon_i^{(1)}\right)\mathbf{X}_{i1}^{(1)} \\
=& \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{X}_{i1}^{(1)}.
\end{aligned}
$$

Since $\widetilde{\mathbf{U}}_1^{(0)} = \mathrm{SVD}_{r_1}(\widetilde{\mathbf{A}}_1)$, the one-sided perturbation bound [19, Proposition 1] yields

$$
\left\|\sin\Theta\left(\widetilde{\mathbf{U}}_1^{(0)}, \mathbf{U}_1\right)\right\| \leq \frac{\sigma_{r_1}(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)\|\mathbf{U}_{1\perp}^\top\widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)^\top}\|}{\sigma_{r_1}^2(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1) - \sigma_{r_1+1}^2(\widetilde{\mathbf{A}}_1)} \wedge 1 \tag{67}
$$

To proceed, we analyze $\sigma_{\min}^2\left(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1\right)$, $\sigma_{r_1+1}(\widetilde{\mathbf{A}}_1)$, and $\|\mathbf{U}_{1\perp}^\top\widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)^\top}\|$, respectively.

- $$
\begin{aligned}
&\sigma_{\min}^2\left(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1\right) \overset{\text{Lemma 2}}{\geq} \sigma_{\min}^2\left(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1\mathbf{W}_1\right) + \sigma_{\min}^2\left(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1(\mathbf{W}_1)_\perp\right) \\
=&\sigma_{\min}^2\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1\right) \\
&+ \sigma_{\min}^2\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}(\mathbf{W}_1)_\perp\right).
\end{aligned}
$$

  By Lemma 4, $\mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1 \in \mathbb{R}^{r_1\times r_1}$, and $n_1 \geq Cp^{3/2}r_1$, we have

  $$
\begin{aligned}
&\sigma_{\min}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1\right) \\
\geq&\sigma_{\min}(\mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1) - \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1 - \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\right\| \\
\overset{\text{Lemma 4}}{\geq}& \sigma_{r_1}(\mathbf{A}_1) - C\sqrt{\frac{\log p}{n_1}\left(2r_1\|\mathbf{A}_1\|_F^2 + \sigma^2\right)} \geq (1-c)\sigma_{r_1}(\mathbf{A}_1)
\end{aligned}
$$

  with probability at least $1-p^{-c}$. When $\mathbf{X}_{i1}^{(1)}$ has i.i.d. Gaussian entries and $\mathbf{W}_1$ is fixed orthogonal matrix, $\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}(\mathbf{W}_1)_\perp \in \mathbb{R}^{r_1\times(p_{-1}-r_1)}$ and $\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i\right) \in \mathbb{R}$ are independently Gaussian distributed and

  $$
\left\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top\mathbf{A}_1\mathbf{W}_1\right\rangle + \varepsilon_i^{(1)} \sim N(0, \widetilde{\sigma}^2).
$$

By Lemma 6,

$$\begin{aligned}
&\sigma_{\min}^2\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top \mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}\mathbf{W}_{1\perp}\right)\\
\geq&\widetilde{\sigma}^2\cdot\frac{n_1 - C_1\sqrt{n_1\log p}}{n_1^2}\cdot\left(\sqrt{p_{-1}-r_1} - \sqrt{r_1} - C_2\sqrt{\log p}\right)^2\\
\geq&\frac{\widetilde{\sigma}^2}{n_1}\cdot\left(1 - C_1\sqrt{\frac{\log p}{n_1}}\right)\cdot\left(p_{-1} - C_3\sqrt{p_{-1}r_1} - C_2\sqrt{p_{-1}\log p}\right)\\
\geq&\frac{\widetilde{\sigma}^2}{n_1}\left(p_{-1} - C_4\sqrt{p_{-1}r_1} - C_5\sqrt{p_{-1}\log p}\right)
\end{aligned}$$

with probability at least $1 - p^{-c}$. To sum up,

$$\sigma_{\min}^2\left(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1\right) \geq (1-c)\sigma_{r_1}^2(\mathbf{A}_1) + \frac{\widetilde{\sigma}^2}{n_1}\cdot\left(p_{-1} - C_1\sqrt{p_{-1}r_1} - C_2\sqrt{p_{-1}\log p}\right) \tag{68}$$

with probability at least $1 - p^{-c}$.

- Next, we consider $\sigma_{r_1+1}(\widetilde{\mathbf{A}}_1)$, note that

$$\begin{aligned}
\sigma_{r_1+1}(\widetilde{\mathbf{A}}_1) = &\min_{\operatorname{rank}(M)\leq r_1}\left\|\widetilde{\mathbf{A}}_1 - \mathbf{M}\right\| \leq \left\|\widetilde{\mathbf{A}}_1 - P_{\mathbf{U}_1}\widetilde{\mathbf{A}}_1\right\| \leq \|\mathbf{U}_{1\perp}^\top\widetilde{\mathbf{A}}_1\|\\
=&\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}, \mathbf{U}_1^\top\mathbf{A}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_{1\perp}^\top\mathbf{X}_{i1}^{(1)}\right\|.
\end{aligned}$$

Since

$$\left(\langle \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top \mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right) \sim N\left(0, \widetilde{\sigma}^2\right),$$

which is also independent of $\mathbf{U}_{1\perp}^\top\mathbf{X}_{i1}^{(1)}$. Thus,

$$\begin{aligned}
\sigma_{r_1+1}^2(\widetilde{\mathbf{A}}_1) =& \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}, \mathbf{U}_1^\top\mathbf{A}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_{1\perp}^\top\mathbf{X}_{i1}^{(1)}\right\|^2\\
\leq&\widetilde{\sigma}^2\cdot\frac{n_1 + C(\sqrt{n_1\log p} + \log p)}{n_1^2}\cdot\left(\sqrt{p_1 - r_1} + \sqrt{p_{-1}} + C\sqrt{\log p}\right)^2\\
\leq&\frac{\widetilde{\sigma}^2}{n_1}\left(1 + C\sqrt{\frac{\log p}{n_1}}\right)\left(p_{-1} + C\sqrt{p_{-1}p_1} + C\sqrt{p_{-1}\log p} + Cp_1 + C\log p\right)\\
\leq&\frac{\widetilde{\sigma}^2}{n_1}\cdot\left(p_{-1} + C\sqrt{p_{-1}p_1} + C\sqrt{p_{-1}\log p} + Cp_1 + C\log p\right)
\end{aligned} \tag{69}$$

with probability at least $1 - p^{-c}$.

- Then we consider $\left\|\mathbf{U}_{1\perp}^\top\widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)^\top}\right\|$. Note that

$$\mathbf{U}_{1\perp}^\top\widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)^\top} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}\mathbf{W}_1, \mathbf{U}_1^\top \mathbf{A}_1\mathbf{W}_1\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_{1\perp}^\top\mathbf{X}_{i1}^{(1)}P_{(\mathbf{U}_1^\top\widetilde{\mathbf{A}}_1)^\top},$$

Here, $\left(\langle \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)} \mathbf{W}_1, \mathbf{U}_1^\top \mathbf{A}_1 \mathbf{W}_1 \rangle + \varepsilon_i\right) \sim N(0, \widetilde{\sigma}^2)$; by independence, conditioning on fixed value of $\mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}$, $\mathbf{U}_{1\perp}^\top \mathbf{X}_{i1}^{(1)}$ is still standard normal, and then

$$\mathbf{U}_{1\perp}^\top \mathbf{X}_{i1}^{(1)} P_{(\mathbf{U}_1^\top \widetilde{\mathbf{A}}_1)^\top} \Big| \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)}$$

is a $(p_1 - r_1)$-by-r_1 i.i.d. standard Gaussian matrix. By Lemma 6, we have

$$\begin{aligned} \left\| \mathbf{U}_{1\perp}^\top \widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top \widetilde{\mathbf{A}}_1)^\top} \right\| \leq & \widetilde{\sigma} \sqrt{\frac{n_1 + C_1\sqrt{n_1 \log p} + C_2 \log p}{n_1^2}} \cdot \left(\sqrt{p_1 - r_1} + \sqrt{r_1} + C_3 \sqrt{\log p} \right) \\ \leq & C_4 \widetilde{\sigma} \cdot \sqrt{\frac{p_1}{n_1}} \end{aligned} \tag{70}$$

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

Combining (68)-(70) with (67), we have the following inequality holds with probability at least $1 - p^{-C}$,

$$\begin{aligned} & \left\| \sin \Theta \left(\widetilde{\mathbf{U}}_1^{(0)}, \mathbf{U}_1 \right) \right\| \\ \leq & \frac{\sigma_{r_1}(\mathbf{U}_1^\top \widetilde{\mathbf{A}}_1) \| \mathbf{U}_{1\perp}^\top \widetilde{\mathbf{A}}_1 P_{(\mathbf{U}_1^\top \widetilde{\mathbf{A}}_1)^\top} \|}{\sigma_{r_1}^2(\mathbf{U}_1^\top \widetilde{\mathbf{A}}_1) - \sigma_{r_1+1}^2(\widetilde{\mathbf{A}}_1)} \wedge 1 \\ \leq & \frac{\left((1-c)\sigma_{r_1}(\mathbf{A}_1) + \widetilde{\sigma}\sqrt{p_{-1}/n_1} \right) \cdot C_1 \widetilde{\sigma} \sqrt{p_1/n_1}}{\left((1-c)\sigma_{r_1}(\mathbf{A}_1) + \widetilde{\sigma}\sqrt{p_{-1}/n_1} \right)^2 - \frac{\widetilde{\sigma}^2}{n_1} \cdot \left(p_{-1} + C_2\sqrt{p_{-1}p_1} + C_3\sqrt{p_{-1}\log p} + C_4 p_1 + C_5 \log p \right)} \wedge 1 \end{aligned}$$

Since $n_1 \geq C p^{3/2} \widetilde{\sigma}^2 / \lambda_0^2$ for large constant $C > 0$, we have

$$\begin{aligned} & \left((1-c)\sigma_{r_1}(\mathbf{A}_1) + \widetilde{\sigma}\sqrt{p_{-1}/n_1} \right)^2 - \frac{\widetilde{\sigma}^2}{n_1} \cdot \left(p_{-1} + C_1\sqrt{p_{-1}p_1} + C_2\sqrt{p_{-1}\log p} + C_3 p_1 + C_4 \log p \right) \\ \geq & (1-c)^2 \sigma_{r_1}^2(\mathbf{A}_1) + 2(1-c)\sigma_{r_1}(\mathbf{A}_1)\widetilde{\sigma}\sqrt{p_{-1}/n_1} - \frac{C_2 \widetilde{\sigma}^2}{n_1} \left(\sqrt{p_1 p_2 p_3} + \sqrt{p_{-1}\log p} + C_3 p_1 + C_4 \log p \right) \\ \geq & c \sigma_{r_1}^2(\mathbf{A}_1) \end{aligned}$$

and additionally,

$$\left\| \sin \Theta \left(\widetilde{\mathbf{U}}_1^{(0)}, \mathbf{U}_1 \right) \right\| \leq \left(\frac{C_1 \widetilde{\sigma} \sqrt{p_1/n_1} \cdot \sigma_{r_1}(\mathbf{A}_1) + \widetilde{\sigma}^2 \sqrt{p_1 p_2 p_3}/n_1}{\sigma_{r_1}^2(\mathbf{A}_1)} \right) \wedge 1.$$

with probability at least $1 - p^{-C}$. Similar inequalities also hold for $\left\| \sin \Theta \left(\widetilde{\mathbf{U}}_2^{(0)}, \mathbf{U}_2 \right) \right\|$ and $\left\| \sin \Theta \left(\widetilde{\mathbf{U}}_3^{(0)}, \mathbf{U}_3 \right) \right\|$. Based on these arguments, we conclude that (66) holds. (66)

further implies that

$$\begin{aligned} e_0 :=& \max_k \left\| \widetilde{\mathbf{U}}_{k\perp}^{(0)\top} \mathcal{M}_k(\boldsymbol{\mathcal{A}}) \right\| = \max_k \left\| \widetilde{\mathbf{U}}_{k\perp}^{(0)\top} \mathbf{U}_k \mathbf{U}_k^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}}) \right\| \\ \leq& \max_k \| \widetilde{\mathbf{U}}_{k\perp}^{(0)\top} \mathbf{U}_k \| \cdot \| \mathbf{U}_k^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}}) \| \leq \max_k \| \sin \Theta(\widetilde{\mathbf{U}}_k^{(0)}, \mathbf{U}_k) \| \cdot \| \mathbf{U}_k^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}}) \| \\ \leq& \max_k C \| \mathbf{A}_k \| \left(\frac{\widetilde{\sigma} \sqrt{p_k / n_1}}{\sigma_{r_k}(\mathbf{A}_k)} + \frac{\widetilde{\sigma}^2 \sqrt{p_1 p_2 p_3} / n_1}{\sigma_{r_k}^2(\mathbf{A}_k)} \right) \\ \leq& C_1 \kappa \left(\frac{\widetilde{\sigma} p^{1/2}}{n_1^{1/2}} + \frac{\widetilde{\sigma}^2 p^{3/2}}{\lambda_0 n_1} \right) \end{aligned} \tag{71}$$

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

Step 2 Then we develop the error bound for $\widetilde{\mathbf{U}}_k$ after enough number of iterations in this step. In particular, we aim to apply Theorem 1 in [138] to give an error bound for the output $\widetilde{\mathbf{U}}_k$ from the high-order order orthogonal iteration (HOOI). To this end, we verify the conditions in Theorem 1 in [138] in this step. Defining

$$\boldsymbol{\mathcal{Z}} = \widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}, \quad \boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{Z}} \times_1 P_{\mathbf{U}_1} \times_2 P_{\mathbf{U}_2} \times_3 P_{\mathbf{U}_3}, \quad \widetilde{\boldsymbol{\mathcal{T}}} = \widetilde{\boldsymbol{\mathcal{A}}}. \tag{72}$$

Then,

$$\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{Z}} \times_1 P_{\mathbf{U}_1} \times_2 P_{\mathbf{U}_2} \times_3 P_{\mathbf{U}_3}. \tag{73}$$

In order to apply Theorem 1 in [138], we develop the following upper bounds under the assumptions of Theorem 4.

- Since $\mathcal{M}_1\left((\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}) \times_1 \mathbf{U}_1^\top \times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top \right)$ is a r_1-by-$(r_2 r_3)$ matrix, Lemma 4 implies

$$\begin{aligned} & \left\| \mathcal{M}_1\left((\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}) \times_1 \mathbf{U}_1^\top \times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top \right) \right\| \\ =& \left\| \mathbf{U}_1^\top \mathcal{M}_1\left(\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}} \right) (\mathbf{U}_3 \otimes \mathbf{U}_2) \right\| \\ =& \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\left\langle \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2), \mathbf{U}_1^\top \mathbf{A}_1 (\mathbf{U}_3 \otimes \mathbf{U}_2) \right\rangle + \varepsilon_i^{(1)} \right) \mathbf{U}_1^\top \mathbf{X}_{i1}^{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2) \right. \\ & \left. - \mathbf{U}_1^\top \mathbf{A}_1 (\mathbf{U}_3 \otimes \mathbf{U}_2) \right\| \\ \overset{\text{Lemma 4}}{\leq}& C_1 \sqrt{\frac{\log p \cdot (r_1 + r_2 r_3) \widetilde{\sigma}^2}{n_1}} \end{aligned} \tag{74}$$

with probability at least $1-p^{-C}$. Similar results also hold for $\mathcal{M}_2(\cdot)$ and $\mathcal{M}_3(\cdot)$. Then

$$\begin{aligned}\lambda_k(\boldsymbol{\mathcal{T}}) :=& \sigma_{r_k}\left(\mathcal{M}_k(\boldsymbol{\mathcal{T}})\right)\\ \overset{(73)}{\geq}& \sigma_{r_k}\left(\mathcal{M}_k(\boldsymbol{\mathcal{A}})\right) - \left\|\mathcal{M}_k\left((\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}) \times_1 P_{\mathbf{U}_1} \times_2 P_{\mathbf{U}_2} \times_3 P_{\mathbf{U}_3}\right)\right\| \\ \geq& \lambda_k - C_1\sqrt{\frac{\log p\cdot(r_k + r_{k+1}r_{k+2})\widetilde{\sigma}^2}{n_1}} \geq (1-c)\lambda_0\end{aligned} \tag{75}$$

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

- Next, we consider

$$\tau_{0k} := \left\|\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\left(\mathbf{U}_{k+2}\otimes\mathbf{U}_{k+1}\right)\right\|, \quad k=1,2,3.$$

In particular,

$$\begin{aligned}&\left\|\mathcal{M}_1(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\left(\mathbf{U}_3\otimes\mathbf{U}_2\right)\right\|\\ \overset{(73)}{=}& \|\mathcal{M}_1(\boldsymbol{\mathcal{Z}} - [\![\boldsymbol{\mathcal{Z}}; P_{\mathbf{U}_1}, P_{\mathbf{U}_2}, P_{\mathbf{U}_3}]\!])(\mathbf{U}_3\otimes\mathbf{U}_2)\|\\ =& \left\|\mathcal{M}_1\left(\left(\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}} - [\![\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}; P_{\mathbf{U}_1}, P_{\mathbf{U}_2}, P_{\mathbf{U}_3}]\!]\right)\times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\right)\right\|\\ =& \Big\|\mathcal{M}_1\left((\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}})\times_1 (P_{\mathbf{U}_1}+P_{\mathbf{U}_{1\perp}})\times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\right)\\ &-\mathcal{M}_1\left((\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}})\times_1 P_{\mathbf{U}_1}\times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\right)\Big\|\\ =& \left\|\mathcal{M}_1\left((\widetilde{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}})\times_1 P_{\mathbf{U}_{1\perp}}\times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\right)\right\|\\ =& \left\|\mathbf{U}_{1\perp}^\top(\widetilde{\mathbf{A}}_1-\mathbf{A}_1)\cdot(\mathbf{U}_3\otimes\mathbf{U}_2)\right\|\\ \leq& \Big\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2), \mathbf{U}_1^\top\mathbf{A}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\rangle + \varepsilon_i^{(1)}\right)\mathbf{U}_{1\perp}^\top\mathbf{X}_{i1}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2)\\ &-\mathbf{U}_{1\perp}^\top\mathbf{A}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\Big\|\\ \overset{\text{Lemma 6}}{\leq}& \widetilde{\sigma}\sqrt{\frac{n_1 + C_1\sqrt{n_1\log p}}{n_1^2}}\left(\sqrt{p_1-r_1}+\sqrt{r_2r_3}+C_2\sqrt{\log p}\right) \leq C_3\widetilde{\sigma}\sqrt{\frac{p_1}{n_1}},\end{aligned} \tag{76}$$

with probability at least $1-p^{-C}$. Thus,

$$\mathbb{P}\left(\tau_{0k}\leq C_1\widetilde{\sigma}\sqrt{p_k/n_1}, \quad k=1,2,3\right) \geq 1-p^{-C}. \tag{77}$$

- Next we consider the upper bound of

$$\begin{aligned}\tau_1 := \max_k\Big\{&\max_{\substack{\mathbf{V}\in\mathbb{R}^{(p_{k+1}-r_{k+1})\times r_{k+1}}\\ \|\mathbf{V}\|\leq 1}}\left\|\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\cdot\{(\mathbf{U}_{k+2,\perp}\mathbf{V})\otimes\mathbf{U}_{k+1}\}\right\|,\\ &\max_{\substack{\mathbf{V}\in\mathbb{R}^{(p_{k+2}-r_{k+2})\times r_{k+2}}\\ \|\mathbf{V}\|\leq 1}}\left\|\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\cdot\{\mathbf{U}_{k+2}\otimes(\mathbf{U}_{k+1,\perp}\mathbf{V})\}\right\|\Big\}.\end{aligned} \tag{78}$$

Note that

$$\begin{aligned}
&\mathcal{M}_1\left(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}}\right)(\mathbf{U}_{3\perp}\mathbf{V})\otimes\mathbf{U}_2\\
=&\left(\mathcal{M}_1(\boldsymbol{\mathcal{Z}})-\mathcal{M}_1\left(\boldsymbol{\mathcal{Z}}\times_1 P_{\mathbf{U}_1}\times_2 P_{\mathbf{U}_2}\times_3 P_{\mathbf{U}_3}\right)\right)(\mathbf{U}_{3\perp}\mathbf{V})\otimes\mathbf{U}_2\\
=&\mathcal{M}_1(\boldsymbol{\mathcal{Z}})(\mathbf{U}_{3\perp}\mathbf{V})\otimes\mathbf{U}_2=\frac{1}{n_1}\sum_{i=1}^{n_1}y_i^{(1)}\mathbf{X}_{i1}^{(1)}((\mathbf{U}_{3\perp}\mathbf{V})\otimes\mathbf{U}_2),
\end{aligned}$$

$$y_i^{(1)}=\langle\boldsymbol{\mathcal{X}}_i^{(1)},\boldsymbol{\mathcal{A}}\rangle+\varepsilon_i^{(1)}=\langle\mathbf{U}_1^\top\mathbf{X}_{i1}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2),\mathbf{U}_1^\top\mathbf{A}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\rangle+\varepsilon_i^{(1)}.$$

Since $\mathbf{U}_{3\perp}$ and $\mathbf{U}_3$ are orthogonal, $y_i^{(1)}$ and $\mathbf{X}_{i1}^{(1)}(\mathbf{U}_{3\perp}\otimes\mathbf{U}_2)$ are independently Gaussian distributed. Thus, conditioning on fixed values of $\{y_i^{(1)}\}_{i=1}^{n_1}$,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}y_i^{(1)}\mathbf{X}_{i1}^{(1)}(\mathbf{U}_{3\perp}\otimes\mathbf{U}_2)\Big|\|\mathbf{y}^{(1)}\|_2^2$$

is a p_1-by-$((p_2-r_2)r_3)$ random matrix with i.i.d. Gaussian entries with mean zero and variance $\|\mathbf{y}^{(1)}\|_2^2/n_1^2$. By Lemma 5 in [139],

$$\begin{aligned}
\mathbb{P}\Bigg(&\max_{\mathbf{V}\in\mathbb{R}^{(p_2-r_2)\times r_2}}\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Z}}(\mathbf{U}_{3\perp}\mathbf{V}\otimes\mathbf{U}_2)\right)\right\|\\
&\geq\frac{C\|\mathbf{y}^{(1)}\|_2}{n_1}\left(\sqrt{p_1}+\sqrt{r_2r_3}+\sqrt{1+t}(\sqrt{p_2r_2}+\sqrt{p_3r_3})\right)\Bigg|\|\mathbf{y}^{(1)}\|_2^2\Bigg)\\
\leq& C\exp\left(-Ct(p_2r_2+p_3r_3)\right).
\end{aligned}\tag{79}$$

Note that $\|\mathbf{y}^{(1)}\|_2^2\sim\widetilde{\sigma}^2\chi_{n_1}^2$, we have

$$\mathbb{P}\left(\|\mathbf{y}^{(1)}\|_2^2\geq\widetilde{\sigma}^2(n_1+2\sqrt{n_1t}+2t)\right)\leq\exp(-t).\tag{80}$$

Combining (79) (with $t=pr/(p_2r_2+p_3r_3)$), (80) (with $t=Cpr$), and the fact that $n_1\geq Cpr$ for large constant $C>0$, we have

$$\mathbb{P}\left(\max_{\substack{\mathbf{V}\in\mathbb{R}^{(p_3-r_2)\times r_1}\\\|\mathbf{V}\|\leq1}}\left\|\mathcal{M}_1\left(\widetilde{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}}\right)(\mathbf{U}_{3\perp}\mathbf{V})\otimes\mathbf{U}_2\right\|\geq C\widetilde{\sigma}\sqrt{\frac{pr}{n_1}}\right)\leq C\exp\left(-cpr\right).$$

By symmetry, we have similar results for other terms in the right hand side of (78) and the following conclusion,

$$\mathbb{P}\left(\tau_1\geq C\widetilde{\sigma}\sqrt{\frac{pr}{n_1}}\right)\leq C\exp(-cpr).\tag{81}$$

- Based on essentially the same argument as the previous step, we can also show

$$\begin{aligned}\tau_2 :=& \max_k \max_{\substack{\mathbf{V}\in\mathbb{R}^{(p_{k+1}-r_{k+1})\times r_{k+1}}:\|\mathbf{V}\|\leq 1;\\ \mathbf{V}'\in\mathbb{R}^{(p_{k+2}-r_{k+2})\times r_{k+2}}:\|\mathbf{V}'\|\leq 1}} \left\|\mathcal{M}_k(\boldsymbol{\mathcal{Z}})\left\{(\mathbf{U}_{k+1\perp}\mathbf{V})\otimes(\mathbf{U}_{k+2\perp}\mathbf{V}')\right\}\right\| \\ \leq& C\widetilde{\sigma}\sqrt{\frac{pr}{n_1}}\end{aligned} \tag{82}$$

  with probability at least $1 - C\exp(-cpr)$.

Now, when the statements in (77), (81), (82) all hold, given $n_1 \geq \frac{\widetilde{\sigma}^2}{\lambda_0^2}(\kappa pr \vee p^{3/2})$ for large enough constant $C > 0$, we have $n_1 \geq \frac{C\widetilde{\sigma}^2}{\lambda_0^2}p^{4/3}r^{1/3}$ (by Hölder's inequality) and the condition

$$\begin{aligned}&\frac{\tau_1}{\lambda(\boldsymbol{\mathcal{T}})} + \max_k \frac{4\tau_2(4\tau_{0k}+e_0)}{\lambda^2(\boldsymbol{\mathcal{T}})} \\ \leq&\frac{C_1\widetilde{\sigma}\sqrt{pr/n_1}}{\lambda_0} + \frac{C_2\widetilde{\sigma}\sqrt{pr/n_1}\left(\widetilde{\sigma}\sqrt{p/n_1}+\kappa\widetilde{\sigma}\sqrt{p/n_1}+\kappa\widetilde{\sigma}^2p^{3/2}/(\lambda_0 n_1)\right)}{\lambda_0^2} \\ \leq&\frac{C_1\widetilde{\sigma}p^{1/2}r^{1/2}}{\lambda_0 n_1^{1/2}} + \frac{C_2\widetilde{\sigma}^2\kappa pr^{1/2}}{\lambda_0^2 n_1} + \frac{C_3\kappa\widetilde{\sigma}^3p^2r^{1/2}}{\lambda_0^3 n_1^{3/2}} \leq 1\end{aligned}$$

holds. Namely, the condition in Theorem 1 in [138] holds when the events of (77), (81), (82) occur.

Step 3 In this step, we try to establish the estimation errors for $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{W}}_k$. First, Theorem 1 in [138] and (77), (81), (82) imply

$$\left\|\sin\Theta\left(\widetilde{\mathbf{U}}_k, \mathbf{U}_k\right)\right\| \leq \frac{C\tau_{0k}}{\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{T}}))} \leq \frac{C\widetilde{\sigma}\sqrt{p_k/n_1}}{\lambda_k}, \quad k = 1,2,3,$$

$$\text{and} \quad \left\|[\![\widetilde{\boldsymbol{\mathcal{T}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{T}}\right\|_{\mathrm{HS}} \leq C\widetilde{\sigma}\sqrt{\frac{p_1r_1+p_2r_2+p_3r_3+r_1r_2r_3}{n_1}}$$

with probability at least $1 - p^{-C}$. Moreover,

$$\begin{aligned}\|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{HS}} \overset{(72)}{=}& \left\|\left(\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right)\times_1 \mathbf{U}_1^\top \times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\right\|_{\mathrm{HS}} \\ =&\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\left\langle\mathrm{vec}(\boldsymbol{\mathcal{X}}_i\times_1\mathbf{U}_1^\top\times_2\mathbf{U}_2^\top\times_3\mathbf{U}_3^\top), \mathrm{vec}(\boldsymbol{\mathcal{A}}\times_1\mathbf{U}_1^\top\times_2\mathbf{U}_2^\top\times_3\mathbf{U}_3^\top)\right\rangle + \varepsilon_i\right)\right. \\ &\left.\cdot\,\mathrm{vec}(\boldsymbol{\mathcal{X}}_i\times_1\mathbf{U}_1^\top\times_2\mathbf{U}_2^\top\times_3\mathbf{U}_3^\top) - \mathrm{vec}(\boldsymbol{\mathcal{A}}\times_1\mathbf{U}_1^\top\times_2\mathbf{U}_2^\top\times_3\mathbf{U}_3^\top)\right\|_2 \\ \overset{\text{Lemma 4}}{\leq}& C\sqrt{\frac{\widetilde{\sigma}^2}{n_1}}\left(\sqrt{r_1r_2r_3}+\sqrt{\log p}\right)\end{aligned}$$

with probability at least $1-p^{-C}$. Combing the previous two inequalities, we have

$$\begin{aligned}
&\left\|[\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}\\
\le& \left\|[\![\widetilde{\boldsymbol{\mathcal{T}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{T}}\right\|_{\mathrm{HS}} + \|\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{T}}\|_{\mathrm{HS}}\\
\le& C\widetilde{\sigma}\sqrt{\frac{p_1r_1+p_2r_2+p_3r_3+r_1r_2r_3}{n_1}} \asymp C\widetilde{\sigma}\sqrt{m/n_1}
\end{aligned} \tag{83}$$

with probability at least $1-p^{-C}$. Then, for $k=1,2,3$,

$$\begin{aligned}
\|\widetilde{\mathbf{U}}_{k\perp}^\top \mathbf{A}_k\|_F \le& \left\|\widetilde{\mathbf{U}}_{k\perp}^\top\left(P_{\widetilde{\mathbf{U}}_k}\widetilde{\mathbf{A}}_k(P_{\widetilde{\mathbf{U}}_{k+2}}\otimes P_{\widetilde{\mathbf{U}}_{k+1}}) - \mathbf{A}_k\right)\right\|_F\\
\le& \left\|P_{\widetilde{\mathbf{U}}_k}\widetilde{\mathbf{A}}_k(P_{\widetilde{\mathbf{U}}_{k+2}}\otimes P_{\widetilde{\mathbf{U}}_{k+1}}) - \mathbf{A}_k\right\| = \left\|[\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}} \le C\widetilde{\sigma}\sqrt{m/n_1}
\end{aligned}$$

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

Next, we are in the position of evaluating the estimation errors of $\widetilde{\mathbf{W}}_k$. Denote $\widetilde{\boldsymbol{\mathcal{S}}} = \widetilde{\boldsymbol{\mathcal{A}}}\times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top$, $\widetilde{\mathbf{V}}_k = \mathrm{SVD}_{r_k}\left(\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right)$, we know

$$\begin{aligned}
\widetilde{\mathbf{W}}_k =& (\widetilde{\mathbf{U}}_{k+2}\otimes\widetilde{\mathbf{U}}_{k+1})\widetilde{\mathbf{V}}_k = \mathrm{SVD}_{r_k}\left((\widetilde{\mathbf{U}}_{k+2}\otimes\widetilde{\mathbf{U}}_{k+1})\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}})^\top\right)\\
=& \mathrm{SVD}_{r_k}\left(\mathcal{M}_k\left(\widetilde{\boldsymbol{\mathcal{S}}}\times_{(k+1)}\widetilde{\mathbf{U}}_{k+1}\times_{(k+2)}\widetilde{\mathbf{U}}_{k+2}\right)^\top\right)\\
=& \mathrm{SVD}_{r_k}\left(\mathcal{M}_k\left(\widetilde{\boldsymbol{\mathcal{S}}}\times_{(k+1)}\widetilde{\mathbf{U}}_{k+1}\times_{(k+2)}\widetilde{\mathbf{U}}_{k+2}\right)^\top\widetilde{\mathbf{U}}_k^\top\right)\\
=& \mathrm{SVD}_{r_k}\left(\mathcal{M}_k\left(\widetilde{\boldsymbol{\mathcal{S}}}\times_k\widetilde{\mathbf{U}}_k\times_{(k+1)}\widetilde{\mathbf{U}}_{k+1}\times_{(k+2)}\widetilde{\mathbf{U}}_{k+2}\right)^\top\right)\\
=& \mathrm{SVD}_{r_k}\left(\mathcal{M}_k\left([\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!]\right)^\top\right).
\end{aligned}$$

On the other hand, $\mathbf{W}_k = \mathrm{SVD}_{r_k}(\mathbf{A}_k^\top) = \mathrm{SVD}_{r_k}\left(\mathcal{M}_k(\boldsymbol{\mathcal{A}})^\top\right)$. By Lemma 7,

$$\begin{aligned}
\|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\|_F \le& 2\left\|\mathcal{M}_k([\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!]) - \mathcal{M}_k(\boldsymbol{\mathcal{A}})\right\|_F\\
=& 2\left\|[\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}} \overset{(83)}{\le} C\widetilde{\sigma}\sqrt{\frac{m}{n_1}}
\end{aligned} \tag{84}$$

with probability at least $1-p^{-C}$. Therefore, we also have

$$\left\|\sin\Theta(\widetilde{\mathbf{W}}_k,\mathbf{W}_k)\right\|_F \le \|\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{W}_k\|_F \le \frac{\|\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{W}_k\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{A}_k^\top\|_F}{\sigma_{r_k}(\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{A}_k^\top)} \le C\sqrt{\frac{\widetilde{\sigma}^2 m}{\lambda_k^2 n_1}} \tag{85}$$

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

To summarize the progress in this step, we have established the following probabilistic inequalities for $\widetilde{\mathbf{U}}_1,\widetilde{\mathbf{U}}_2,\widetilde{\mathbf{U}}_3$ and $\widetilde{\mathbf{W}}_1,\widetilde{\mathbf{W}}_2,\widetilde{\mathbf{W}}_3$,

$$\left\|\sin\Theta\left(\widetilde{\mathbf{U}}_k,\mathbf{U}_k\right)\right\| \le \frac{C\widetilde{\sigma}\sqrt{p_k/n_1}}{\lambda_k}, \quad \left\|\sin\Theta\left(\widetilde{\mathbf{W}}_k,\mathbf{W}_k\right)\right\|_F \le \frac{C\widetilde{\sigma}\sqrt{m/n_1}}{\lambda_k}, \quad k=1,2,3, \tag{86}$$

$$\left\|\widetilde{\mathbf{U}}_k^\top \mathbf{A}_k\right\|_F \leq C\widetilde{\sigma}\sqrt{m/n_1}, \quad \left\|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\right\|_F \leq C\widetilde{\sigma}\sqrt{m/n_1}, \quad k=1,2,3, \tag{87}$$

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

Step 4 For the rest of the proof, we assume (86) and (87) hold. Next, we move on to evaluate the estimation error bound for $\widehat{\boldsymbol{\mathcal{A}}}$. The focus now shifts from the first batch of samples $(\boldsymbol{\mathcal{X}}^{(1)}, \mathbf{y}^{(1)})$ to the second one $(\boldsymbol{\mathcal{X}}^{(2)}, y^{(2)})$. Denote

$$\theta_k := \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_k, \mathbf{U}_k\right)\right\| \overset{(86)}{\leq} \frac{C\widetilde{\sigma}\sqrt{p_k/n_1}}{\lambda_k}, \quad k=1,2,3; \tag{88}$$

$$\xi_k := \|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\|_F \overset{(87)}{\leq} C\widetilde{\sigma}\sqrt{m/n_1}, \quad k=1,2,3; \tag{89}$$

$$\eta_k := \left\|\widetilde{\mathbf{U}}_k^\top \mathbf{A}_k\right\|_F \overset{(87)}{\leq} C\widetilde{\sigma}\sqrt{m/n_1}, \quad k=1,2,3; \tag{90}$$

$$\widehat{\sigma}^2 := \left\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2^2 + \sigma^2. \tag{91}$$

By Lemma 9,

$$\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \leq \frac{C\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2} + \frac{C_1\widetilde{\sigma}^6 mp^2}{\lambda_0^4 n_1^3}.$$

Provided that $m = r_1r_2r_3 + \sum_k(p_k-r_k)r_k$ and $n_1 \geq \frac{C\widetilde{\sigma}^2 p}{\lambda_0^2}$, we know

$$\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \leq \frac{C\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}, \quad \widehat{\sigma}^2 \leq \sigma^2 + \frac{C\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}. \tag{92}$$

Step 5 In this step, we evaluate two crucial quantities for applying the oracle inequality (Theorem 2). Recall the importance sketching covariates (6) are defined as

$$\widetilde{\mathbf{X}} = \left[\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_1} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_2} \;\; \widetilde{\mathbf{X}}_{\mathbf{D}_3}\right] \in \mathbb{R}^{n_2\times m},$$
$$\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n\times(r_1r_2r_3)}, \quad \left(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\right)_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i^{(2)} \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_2^\top \times_3 \widetilde{\mathbf{U}}_3^\top\right),$$
$$\widetilde{\mathbf{X}}_{\mathbf{D}_k} \in \mathbb{R}^{n\times(p_k-r_k)r_k}, \quad \left(\widetilde{\mathbf{X}}_{\mathbf{D}_k}\right)_{[i,:]} = \mathrm{vec}\left(\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i^{(2)} \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right)\widetilde{\mathbf{V}}_k\right).$$

When $\boldsymbol{\mathcal{X}}_i^{(2)}$ are i.i.d. Gaussian matrices and independent of $\widetilde{\mathbf{U}}_k$, $\widetilde{\mathbf{V}}_k$, $\widetilde{\mathbf{W}}_k$, $\widetilde{\mathbf{X}}$ can be seen as an orthogonal projection of $\boldsymbol{\mathcal{X}}_i^{(2)}$ and has i.i.d. Gaussian entries. Thus, by Proposition 5.35 in [122],

$$\mathbb{P}\left(\sigma_{\min}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}) = \sigma_{\min}^2(\widetilde{\mathbf{X}}) \geq \left(\sqrt{n_2}-\sqrt{m}-t\right)^2\right) \geq 1-\exp(-t^2/2).$$

By definition, $\widetilde{\varepsilon} \in \mathbb{R}^{n_2}$ is independent of $\widetilde{\mathbf{X}}$, and

$$\widetilde{\varepsilon}_j = \langle \boldsymbol{\mathcal{X}}_j^{(2)}, P_{\widetilde{\mathbf{U}}_\perp}\boldsymbol{\mathcal{A}}\rangle + \varepsilon_j \sim N\left(0, \left\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2^2 + \sigma^2\right) = N(0, \widehat{\sigma}^2).$$

Then, $\|\widetilde{\boldsymbol{\varepsilon}}\|_2^2 \sim \widehat{\sigma}^2\chi^2_{n_2}$ and $\|\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{\varepsilon}}\|_2^2\Big|\|\boldsymbol{\varepsilon}\|_2^2 \sim \|\boldsymbol{\varepsilon}\|_2^2\chi^2_m$. Based on χ^2 distribution tail bound [72, Lemma 1] and $n_2 \geq C(p^{3/2}+r^3) \geq Cm$,

$$\begin{aligned}
&\left\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{\varepsilon}}\right\|_2^2\\
\leq&\frac{\widehat{\sigma}^2\left(n_2+2\sqrt{n_2C_1\log(p)}+2C_2\log(p)\right)\left(m+2\sqrt{mC_3\log(p)}+2C\log(p)\right)}{\left(\sqrt{n_2}-\sqrt{m}-C_4\log(p)\right)^4}\\
\leq&\frac{\widehat{\sigma}^2 m}{n_2}\frac{\left(1+2\sqrt{\frac{C\log p}{n_2}}+2\frac{\log p}{n_2}\right)\left(1+2\sqrt{\frac{t}{m}}+2\frac{t}{m}\right)}{\left(1-\sqrt{\frac{m}{n_2}}-\frac{C_1\log(p)}{\sqrt{n_2}}\right)^4}\\
=&\frac{\widehat{\sigma}^2 m}{n_2}\left(1+C_1\sqrt{\frac{m}{n_2}}+C_2\sqrt{\frac{\log p}{m}}\right).
\end{aligned}\tag{93}$$

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

We assume (93) holds. It remains to check $\left\|\widehat{\mathbf{D}}_k(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\|$. Similarly as the proof of Theorem 2, we define

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{B}}}=&\left[\!\left[\boldsymbol{\mathcal{A}};\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top\right]\!\right]=\left[\!\left[\boldsymbol{\mathcal{S}}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times_3\mathbf{U}_3;\widetilde{\mathbf{U}}_1^\top,\widetilde{\mathbf{U}}_2^\top,\widetilde{\mathbf{U}}_3^\top\right]\!\right]\in\mathbb{R}^{r_1\times r_2\times r_3};\\
\widetilde{\mathbf{B}}_k=&\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{B}}})\in\mathbb{R}^{r_k\times(r_{k+1}r_{k+2})},\quad k=1,2,3,\\
\widetilde{\mathbf{D}}_1=&\widetilde{\mathbf{U}}_{1\perp}^\top\mathcal{M}_1(\boldsymbol{\mathcal{A}}\times_2\widetilde{\mathbf{U}}_2^\top\times_3\widetilde{\mathbf{U}}_3)\widetilde{\mathbf{V}}_1\overset{\text{Lemma 1}}{=}\widetilde{\mathbf{U}}_{1\perp}^\top\mathbf{A}_1\widetilde{\mathbf{W}}_1\in\mathbb{R}^{(p_1-r_1)\times r_1},\\
\widetilde{\mathbf{D}}_2=&\widetilde{\mathbf{U}}_{2\perp}^\top\mathcal{M}_2(\boldsymbol{\mathcal{A}}\times_1\widetilde{\mathbf{U}}_1^\top\times_3\widetilde{\mathbf{U}}_3)\widetilde{\mathbf{V}}_2=\widetilde{\mathbf{U}}_{2\perp}^\top\mathbf{A}_2\widetilde{\mathbf{W}}_2\in\mathbb{R}^{(p_2-r_2)\times r_2},\\
\widetilde{\mathbf{D}}_3=&\widetilde{\mathbf{U}}_{3\perp}^\top\mathcal{M}_3(\boldsymbol{\mathcal{A}}\times_1\widetilde{\mathbf{U}}_1^\top\times_2\widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_3=\widetilde{\mathbf{U}}_{3\perp}^\top\mathbf{A}_3\widetilde{\mathbf{W}}_3\in\mathbb{R}^{(p_3-r_3)\times r_3}.
\end{aligned}$$

By the proof of Theorem 2, we have

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\mathcal{B}}}-\widetilde{\boldsymbol{\mathcal{B}}}\right\|_{\text{HS}}^2+\sum_{k=1}^3\left\|\widehat{\mathbf{D}}_k-\widetilde{\mathbf{D}}_k\right\|_F^2\overset{(41)}{\leq}\left\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\varepsilon}\right\|_2^2\\
&\overset{(93)}{\leq}\frac{\widehat{\sigma}^2 m}{n_2}\left(1+C_1 s\sqrt{\frac{\log m}{n_2}}+C_2\sqrt{\frac{\log p}{m}}\right),
\end{aligned}\tag{94}$$

$$\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\|\overset{(49)}{\leq}C\max_k\left\{\|\sin\Theta(\widetilde{\mathbf{U}}_k,\mathbf{U}_k)\|,\|\sin\Theta(\mathbf{W}_k,\mathbf{W}_k)\|\right\}\leq\frac{C\widetilde{\sigma}\sqrt{m/n_1}}{\lambda_k},\tag{95}$$

$$\sigma_{\min}(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)=\sigma_{\min}(\widetilde{\mathbf{U}}_k^\top\mathbf{A}_k\widetilde{\mathbf{W}}_k)\overset{(42)}{\geq}\lambda_k\left(1-\frac{C\widetilde{\sigma}^2 m}{\lambda_k^2 n_1}\right)\geq\lambda_k(1-c)$$

for some constant $0<c<1$. This additionally means

$$\sigma_{\min}\left(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k\right)\geq\sigma_{\min}(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)-\|\widehat{\mathbf{B}}_k-\mathbf{B}_k\|\overset{(94)}{\geq}\lambda_k\left(1-\frac{C\widetilde{\sigma}^2 m}{\lambda_k^2 n_1}\right)-\frac{C\widehat{\sigma}^2 m}{n_2}\geq(1-c)\lambda_k.\tag{96}$$

It is easy to check that the following equality,

$$(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1} = (\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1} + (\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\left(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k - \widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k\right)(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}.$$

Thus,

$$\begin{aligned}\rho := \left\|\widehat{\mathbf{D}}_k(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\| \leq & \left\|(\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k)(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\| + \left\|\widetilde{\mathbf{D}}_k(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\| \\ \leq & \frac{C\left\|\widehat{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k\right\|}{\lambda_k} + \left\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\| \\ & + \left\|\widetilde{\mathbf{D}}_k(\widetilde{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\right\| \cdot \left\|(\widetilde{\mathbf{B}}_k - \widehat{\mathbf{B}}_k)\widetilde{\mathbf{V}}_k\right\| \cdot \|(\widehat{\mathbf{B}}_k\widetilde{\mathbf{V}}_k)^{-1}\| \\ \overset{(94)(95)(96)}{\leq} & \frac{C\widetilde{\sigma}}{\lambda_k}\sqrt{\frac{m}{n_1}} + \frac{C\widehat{\sigma}}{\lambda_k}\sqrt{\frac{m}{n_2}}.\end{aligned} \tag{97}$$

Step 6 Finally, we apply the oracle inequality, i.e., Theorem 2, and obtain the final upper bound for $\widehat{\mathcal{A}}$. We have shown that the conditions of Theorem 2 holds if (86), (87), and (93) hold. Then Theorem 2 implies

$$\begin{aligned}\left\|\widehat{\mathcal{A}} - \mathcal{A}\right\|_{\mathrm{HS}}^2 \leq & (1 + C\theta + C\rho)\left\|(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\varepsilon}\right\|_2^2 \\ \overset{(88)(93)(97)}{\leq} & \frac{\widehat{\sigma}m}{n_2}\left(1 + C_1\sqrt{\frac{m}{n_2}} + C_2\sqrt{\frac{\log p}{m}} + \frac{C_3\widetilde{\sigma}}{\lambda_0}\sqrt{\frac{m}{n_1}} + \frac{C_4\widehat{\sigma}}{\lambda_0}\sqrt{\frac{m}{n_2}}\right) \\ \overset{(92)}{\leq} & \frac{m}{n_2}\left(\sigma^2 + \frac{C_1\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}\right)\left(1 + C_2\sqrt{\frac{m}{n_2}} + C_3\sqrt{\frac{\log p}{m}} + \frac{C_4\widetilde{\sigma}}{\lambda_0}\sqrt{\frac{m}{n_1}} + \frac{C_5\widehat{\sigma}}{\lambda_0}\sqrt{\frac{m}{n_2}}\right) \\ \leq & \frac{m}{n_2}\left(\sigma^2 + \frac{C_1\widetilde{\sigma}^4 mp}{n_1^2\lambda_0^2}\right)\left(1 + C_2\sqrt{\frac{\log p}{m}} + C_3\sqrt{\frac{m\widetilde{\sigma}^2}{(n_1 \wedge n_2)\lambda_0^2}}\right)\end{aligned}$$

with probability at least $1 - p^{-C}$. Here, the last inequality is due to $n_1 \wedge n_2 \geq C\widetilde{\sigma}^2(p^{3/2} + r^3)/\lambda_0^2$ and $\widehat{\sigma} = \|\mathcal{A}\|_{\mathrm{HS}}^2 + \sigma^2 \geq \lambda_0$. □

F.4 Proof of Theorem 5

In this theorem, we provide an estimation error lower bound for low-rank tensor regression. The central idea is to carefully transform the original high-dimensional low-rank tensor regression model to the unconstrained dimension-reduced linear regression model (103), then apply the classic Bayes risk of linear regression (Lemma 10) to finalize the desired lower bound on estimation error.

Since r_1, r_2, and r_3 satisfy $r_k \leq r_{k+1}r_{k+2}$ for $k = 1, 2, 3$, the r_1-by-r_2-by-r_3 tensor with i.i.d. normal entries has full Tucker rank with probability 1. Thus, we can set $\mathcal{S}_0 \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ as a fixed tensor with full Tucker rank, i.e., $\mathrm{rank}(\mathcal{S}_0) = (r_1, r_2, r_3)$. Let $T > 0$ be a large to-be-specified constant. Define

$$\mathcal{A}_0 \in \mathbb{R}^{p_1 \times p_2 \times p_3}, \quad (\mathcal{A}_0)_{[1:r_1, 1:r_2, 1:r_3]} = T\mathcal{S}_0, \quad (\mathcal{A}_0)_{[1:r_1, 1:r_2, 1:r_3]^c} = 0. \tag{98}$$

Suppose $\mathbf{U}_k \in \mathbb{O}_{p_k,r_k}$ and $\mathbf{W}_k \in \mathbb{O}_{p_{-k},r_k}$ are the left and right singular subspaces of $\mathcal{M}_k(\boldsymbol{\mathcal{A}}_0)$, respectively; $\mathbf{V}_k \in \mathbb{O}_{r_{k+1}r_{k+2},r_k}$ is the right singular subspace of $\mathcal{M}_k(\boldsymbol{\mathcal{S}}_0)$. Then by definition of $\boldsymbol{\mathcal{A}}_0$,

$$\mathbf{U}_k = \begin{bmatrix} \mathbf{I}_{r_k} \\ \mathbf{0}_{(p_k-r_k)\times r_k} \end{bmatrix}, \quad k = 1, 2, 3.$$

Next, for to-be-specified values $\tau, T > 0$, we introduce a prior distribution $\bar{P}_{\tau,T}$ on the class of $\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}$: the p_1-by-p_2-by-p_3 random tensor $\bar{\boldsymbol{\mathcal{A}}} \sim \bar{P}_{\tau,T}$ if and only if it can be generated based on the following process.

1. Generate an r_1-by-r_2-by-r_3 tensor $\boldsymbol{\mathcal{B}} \overset{iid}{\sim} N(0,\tau^2)$ and assign $\bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]} = T\boldsymbol{\mathcal{S}}_0 + \boldsymbol{\mathcal{B}}$.

2. Suppose $\mathcal{M}_k(\bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]}) = \bar{\mathbf{A}}_{0k} \in \mathbb{R}^{r_k\times r_{-k}}$ and $\bar{\mathbf{V}}_k = \text{SVD}_{r_k}(\bar{\mathbf{A}}_{0k}^\top) \in \mathbb{O}_{r_{-k},r_k}$. Assign
$$\begin{aligned} \mathcal{M}_1\left(\bar{\boldsymbol{\mathcal{A}}}_{[(r_1+1):p_1,1:r_2,1:r_3]}\right) &= \mathbf{B}_1 \cdot \bar{\mathbf{V}}_1^\top, \\ \mathcal{M}_2\left(\bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,(r_2+1):p_2,1:r_3]}\right) &= \mathbf{B}_2 \cdot \bar{\mathbf{V}}_2^\top, \\ \mathcal{M}_3\left(\bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,(r_3+1):p_3]}\right) &= \mathbf{B}_3 \cdot \bar{\mathbf{V}}_3^\top, \end{aligned}$$
where all entries of $\mathbf{B}_1 \in \mathbb{R}^{(p_1-r_1)\times r_1}, \mathbf{B}_2 \in \mathbb{R}^{(p_2-r_2)\times r_2}, \mathbf{B}_3 \in \mathbb{R}^{(p_3-r_3)\times r_3}$ are independently drawn from $N(0,\tau^2)$.

3. The other blocks of $\bar{\boldsymbol{\mathcal{A}}}$ are calculated as follows,
$$\begin{aligned} &\bar{\boldsymbol{\mathcal{A}}}_{[(r_1+1):p_1,(r_2+1):p_2,1:r_3]} = \bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]} \times_1 \left(\mathbf{B}_1(\bar{\mathbf{A}}_{01}\bar{\mathbf{V}}_1)^{-1}\right) \times_2 \left(\mathbf{B}_2(\bar{\mathbf{A}}_{02}\bar{\mathbf{V}}_2)^{-1}\right), \\ &\bar{\boldsymbol{\mathcal{A}}}_{[(r_1+1):p_1,1:r_2,(r_3+1):p_3]} = \bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]} \times_1 \left(\mathbf{B}_1(\bar{\mathbf{A}}_{01}\bar{\mathbf{V}}_1)^{-1}\right) \times_3 \left(\mathbf{B}_3(\bar{\mathbf{A}}_{03}\bar{\mathbf{V}}_3)^{-1}\right), \\ &\bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,(r_2+1):p_2,(r_3+1):p_3]} = \bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]} \times_2 \left(\mathbf{B}_2(\bar{\mathbf{A}}_{02}\bar{\mathbf{V}}_2)^{-1}\right) \times_3 \left(\mathbf{B}_3(\bar{\mathbf{A}}_{03}\bar{\mathbf{V}}_3)^{-1}\right), \qquad (99) \\ &\bar{\boldsymbol{\mathcal{A}}}_{[(r_1+1):p_1,(r_2+1):p_2,(r_3+1):p_3]} \\ &\quad = \bar{\boldsymbol{\mathcal{A}}}_{[1:r_1,1:r_2,1:r_3]} \times_1 \left(\mathbf{B}_1(\bar{\mathbf{A}}_{01}\bar{\mathbf{V}}_1)^{-1}\right) \times_2 \left(\mathbf{B}_2(\bar{\mathbf{A}}_{02}\bar{\mathbf{V}}_2)^{-1}\right) \times_3 \left(\mathbf{B}_3(\bar{\mathbf{A}}_{03}\bar{\mathbf{V}}_3)^{-1}\right). \end{aligned}$$

One can check by comparing each block that $\bar{\boldsymbol{\mathcal{A}}}$ satisfies

$$\bar{\boldsymbol{\mathcal{A}}} = \llbracket T\boldsymbol{\mathcal{S}}_0 + \boldsymbol{\mathcal{B}}; \bar{\mathbf{L}}_1, \bar{\mathbf{L}}_2, \bar{\mathbf{L}}_3 \rrbracket, \quad \text{where} \quad \bar{\mathbf{L}}_k = \begin{bmatrix} \mathbf{I}_{r_k} \\ \mathbf{B}_k(\bar{\mathbf{A}}_{0k}\bar{\mathbf{V}}_k)^{-1} \end{bmatrix}, \quad k = 1, 2, 3. \qquad (100)$$

Thus, $\text{rank}(\bar{\boldsymbol{\mathcal{A}}}) \leq (r_1, r_2, r_3)$ and $\bar{\boldsymbol{\mathcal{A}}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}$. Then we consider another distribution $P^*_{\tau,T}$ on the whole tensor space $\mathbb{R}^{p_1\times p_2\times p_3}$,

$$\boldsymbol{\mathcal{A}}^* \sim P^*_{\tau,T}, \quad \text{such that} \quad \begin{aligned} &\boldsymbol{\mathcal{A}}^*_{[1:r_1,1:r_2,1:r_3]} = T\boldsymbol{\mathcal{S}}_0 + \boldsymbol{\mathcal{B}}, \\ &\mathcal{M}_1\left(\boldsymbol{\mathcal{A}}^*_{[(r_1+1):p_1,1:r_2,1:r_3]}\right) = \mathbf{B}_1 \cdot \mathbf{V}_1^\top; \\ &\mathcal{M}_2\left(\boldsymbol{\mathcal{A}}^*_{[1:r_1,(r_2+1):p_2,1:r_3]}\right) = \mathbf{B}_2 \cdot \mathbf{V}_2^\top; \\ &\mathcal{M}_3\left(\boldsymbol{\mathcal{A}}^*_{[(r_1+1):p_1,1:r_2,1:r_3]}\right) = \mathbf{B}_3 \cdot \mathbf{V}_3^\top; \\ &\text{the other blocks of } \boldsymbol{\mathcal{A}}^* \text{ are set to zero.} \end{aligned} \qquad (101)$$

Here, $\boldsymbol{\mathcal{B}}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3 \overset{iid}{\sim} N(0, \tau^2)$. Suppose $\bar{\boldsymbol{\mathcal{A}}} \sim \bar{P}_{\tau,T}$ and $\boldsymbol{\mathcal{A}}^* \sim P^*_{\tau,T}$. Recall that $\mathbf{V}_k = \mathrm{SVD}_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{S}}_0)^\top))$ and $\bar{\mathbf{V}}_k = \mathrm{SVD}_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{S}}_0 + \boldsymbol{\mathcal{B}}/T)^\top)$. As $T \to \infty$, we must have

$$\bar{\mathbf{V}}_k \xrightarrow{d} \mathbf{V}_k \quad \text{and} \quad (\bar{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}_0) \xrightarrow{d} (\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}_0). \tag{102}$$

Next, we move on to the regular tensor regression model

$$y_i = \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}} \rangle + \varepsilon_i, \quad i = 1, \ldots, n.$$

For convenience, we divide $\boldsymbol{\mathcal{X}}_i$ and $\boldsymbol{\mathcal{A}}$ into eight blocks and denote them separately as

$$\boldsymbol{\mathcal{X}}_{i,s_1s_2s_3} = (\boldsymbol{\mathcal{X}}_i)_{[I_{1,s_1}, I_{2,s_2}, I_{3,s_3}]}, \quad \boldsymbol{\mathcal{A}}_{s_1s_2s_3} = \boldsymbol{\mathcal{A}}_{[I_{1,s_1}, I_{2,s_2}, I_{3,s_3}]}, \text{ for } s_1, s_2, s_3 \in \{1, 2\},$$
$$\text{where} \quad I_{k,1} = \{1, \ldots, r_k\}, \quad I_{k,2} = \{r_k + 1, \ldots, p_k\}, \quad k = 1, 2, 3.$$

If $\boldsymbol{\mathcal{A}}^* \sim P^*_{\tau,T}$, $\boldsymbol{\mathcal{A}}^*_{122}, \boldsymbol{\mathcal{A}}^*_{212}, \boldsymbol{\mathcal{A}}^*_{221}, \boldsymbol{\mathcal{A}}^*_{222}$ are all zeros. Then,

$$\begin{aligned}
y_i =& \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}}^* \rangle + \varepsilon_i = \sum_{s_1,s_2,s_3=1}^{2} \langle \boldsymbol{\mathcal{X}}_{i,s_1s_2s_3}, \boldsymbol{\mathcal{A}}^*_{s_1s_2s_3} \rangle + \varepsilon_i \\
=& \langle (\boldsymbol{\mathcal{X}}_{i,111}, T\boldsymbol{\mathcal{S}}_0 + \boldsymbol{\mathcal{B}} \rangle + \langle \mathcal{M}_1(\boldsymbol{\mathcal{X}}_{i,211}), \mathbf{B}_1 \mathbf{V}_1^\top \rangle \\
& + \langle \mathcal{M}_2(\boldsymbol{\mathcal{X}}_{i,121}), \mathbf{B}_2 \mathbf{V}_2^\top \rangle + \langle \mathcal{M}_3(\boldsymbol{\mathcal{X}}_{i,112}), \mathbf{B}_3 \mathbf{V}_3^\top \rangle + \varepsilon_i \\
=& \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}}_0 \rangle + \varepsilon_i + \langle \mathrm{vec}(\boldsymbol{\mathcal{X}}_{i,111}), \mathrm{vec}(\boldsymbol{\mathcal{B}}) \rangle + \langle \mathcal{M}_1(\boldsymbol{\mathcal{X}}_{i,211}) \mathbf{V}_1, \mathbf{B}_1 \rangle \\
& + \langle \mathcal{M}_2(\boldsymbol{\mathcal{X}}_{i,121}) \mathbf{V}_2, \mathbf{B}_2 \rangle + \langle \mathcal{M}_3(\boldsymbol{\mathcal{X}}_{i,112}) \mathbf{V}_3, \mathbf{B}_3 \rangle \\
:=& \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}}_0 \rangle + \langle \bar{\mathbf{X}}_i, \mathbf{b} \rangle + \varepsilon_i,
\end{aligned}$$

where

$$\bar{\mathbf{X}}_i = \begin{bmatrix} \mathrm{vec}\,(\boldsymbol{\mathcal{X}}_{i,111}) \\ \mathrm{vec}\,(\mathcal{M}_1(\mathbf{X}_{i,211})\mathbf{V}_1) \\ \mathrm{vec}\,(\mathcal{M}_2(\mathbf{X}_{i,121})\mathbf{V}_2) \\ \mathrm{vec}\,(\mathcal{M}_3(\mathbf{X}_{i,112})\mathbf{V}_3) \end{bmatrix} \in \mathbb{R}^m, \quad \bar{\mathbf{X}} = \begin{bmatrix} \bar{\mathbf{X}}_1^\top \\ \vdots \\ \bar{\mathbf{X}}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad \mathbf{b} = \begin{bmatrix} \mathrm{vec}(\boldsymbol{\mathcal{B}}) \\ \mathrm{vec}(\mathbf{B}_1) \\ \mathrm{vec}(\mathbf{B}_2) \\ \mathrm{vec}(\mathbf{B}_3) \end{bmatrix} \in \mathbb{R}^m.$$

Suppose the parameter $\boldsymbol{\mathcal{A}}^*$ is drawn from the prior distribution $P^*_{\tau,T}$. Then, $\mathbf{b} \overset{iid}{\sim} N(0, \tau^2)$. Note that $\bar{\mathbf{X}}_i$ is an orthogonal projection of $\boldsymbol{\mathcal{X}}_i$, so $\bar{\mathbf{X}}_i \overset{iid}{\sim} N(0, 1)$. Now, $y_i, \bar{\mathbf{X}}_i, \bar{\mathbf{b}}$ can be related by the following regression model,

$$\begin{aligned}
& y_i - \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{A}}_0 \rangle = \bar{\mathbf{X}}_i^\top \mathbf{b} + \varepsilon_i, \quad i = 1, \ldots, n; \\
& \mathbf{b} \overset{iid}{\sim} N(0, \tau^2), \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2).
\end{aligned} \tag{103}$$

By the construction of $\boldsymbol{\mathcal{A}}^*$ and the setting that $\boldsymbol{\mathcal{S}}_0$ is fixed, the estimation of $\boldsymbol{\mathcal{A}}^*$ is equivalent to the estimation $\mathbf{b}$. By Lemma 10, the Bayes risk of estimating $\mathbf{b}$ (and the Bayes risk of estimating $\boldsymbol{\mathcal{A}}^*$ if $\boldsymbol{\mathcal{A}}^* \sim P_{\tau,T}$) is

$$\left\| \widehat{\boldsymbol{\mathcal{A}}}^* - \boldsymbol{\mathcal{A}}^* \right\|_{\mathrm{HS}}^2 \Big| \{\bar{\mathbf{X}}_i\}_{i=1}^n = \left\| \widehat{\mathbf{b}} - \mathbf{b} \right\|_2^2 \Big| \{\bar{\mathbf{X}}_i\}_{i=1}^n = \mathrm{tr}\left(\left(\frac{\mathbf{I}_m}{\tau^2} + \frac{\bar{\mathbf{X}}^\top \bar{\mathbf{X}}}{\sigma^2} \right)^{-1} \right).$$

Here, $\widehat{\boldsymbol{\mathcal{A}}}^*$ and $\widehat{\mathbf{b}}$ are the posterior mean of $\boldsymbol{\mathcal{A}}^*$ and $\mathbf{b}$, respectively.

Since $\bar{P}_{\tau,T} \to P_{\tau,T}$ and $\bar{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}_0 \to \boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}_0$ as $T \to \infty$, we have

$$\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \bar{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{HS}}^2 \Big|\{\bar{\mathbf{X}}_i\}_{i=1}^n \to \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}}^* - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{HS}}^2 \Big|\{\bar{\mathbf{X}}_i\}_{i=1}^n = \mathrm{tr}\left(\left(\frac{\mathbf{I}_m}{\tau^2} + \frac{\bar{\mathbf{X}}^\top \bar{\mathbf{X}}}{\sigma^2}\right)^{-1}\right),$$

where $\widehat{\boldsymbol{\mathcal{A}}}$ is the posterior mean of $\bar{\boldsymbol{\mathcal{A}}}$ if $\bar{\boldsymbol{\mathcal{A}}} \sim \bar{P}_{\tau,T}$. Since $\bar{\boldsymbol{\mathcal{A}}} \sim \bar{P}_{\tau,T}$ and $\bar{P}_{\tau,T}$ is the distribution on $\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}$, we have the following estimation lower bound,

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \Big|\{\bar{\mathbf{X}}_i\}_{i=1}^n \geq \mathrm{tr}\left(\left(\frac{\mathbf{I}_m}{\tau^2} + \frac{\bar{\mathbf{X}}^\top \bar{\mathbf{X}}}{\sigma^2}\right)^{-1}\right).$$

Finally, since $(\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^{-1}$ is inverse Wishart distributed and[6]

$$\mathrm{tr}(\mathbb{E}(\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^{-1}) = \begin{cases} \frac{1}{n-m-1}\mathrm{tr}(\mathbf{I}_m) = \frac{m}{n-m-1} & n > m+1; \\ \infty & n \leq m+1. \end{cases}$$

By letting $\tau \to \infty$, we finally obtain

$$\begin{aligned} &\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \limsup_{\tau \to \infty} \mathbb{E}\mathrm{tr}\left(\left(\frac{\mathbf{I}_m}{\tau^2} + \frac{\bar{\mathbf{X}}^\top \bar{\mathbf{X}}}{\sigma^2}\right)^{-1}\right) \\ =&\mathrm{tr}\left(\frac{\sigma^2 \mathbf{I}_m}{n-m-1}\right) = \begin{cases} \frac{m\sigma^2}{n-m-1}, & \text{if } n > m+1; \\ +\infty & \text{if } n \leq m+1. \end{cases} \end{aligned}$$

□

F.5 Proof of Theorem 6

In this theorem, we aim to establish an estimation error upper bound for sparse ISLET in sparse low-rank tensor regression problem. After introducing some necessary notations, we develop the estimation error bounds for sketching directions $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{W}}_k$ in Steps 1 and 2. In Step 3, we give error bounds for a number of intermediate terms. In Step 4, we prove upper bounds for key quantities ρ, $\left\|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1} \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{B}}\right\|_2^2$, $\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1} \widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2$, and $\max_{i=1,\ldots,p_k} \left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k,[:,G_i^k]})^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n\right\|_2^2$. Finally, we plug in these values to Theorem 3 to finalize the proof.

We first introduce a number of notations that will be used in the proof. Similarly as the proof of Theorem 4, denote

$$\mathbf{A}_k = \mathcal{M}_k(\boldsymbol{\mathcal{A}}), \quad \mathbf{S}_k = \mathcal{M}_k(\boldsymbol{\mathcal{S}}),$$

[6] See `https://en.wikipedia.org/wiki/Inverse-Wishart_distribution` for expectation of inverse Wishart distribution.

$$\widetilde{\mathbf{A}}_k = \mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{A}}}), \quad \widetilde{\boldsymbol{\mathcal{S}}} = [\![\widetilde{\boldsymbol{\mathcal{A}}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!], \quad \widetilde{\mathbf{S}}_k = \mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{S}}}), \quad \mathbf{X}_{jk} = \mathcal{M}_k(\boldsymbol{\mathcal{X}}_j), \quad k = 1,2,3.$$

Recall

$$\widetilde{\sigma}^2 = \|\boldsymbol{\mathcal{A}}\|_{\text{HS}}^2 + \sigma^2, \quad \lambda_k = \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}})),$$
$$m_s = r_1r_2r_3 + \sum_{k\in J_s} s_k(r_k + \log(p_k)) + \sum_{k\notin J_s} p_k r_k, \tag{104}$$

and $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_2$ are the output from Step 1. We also denote

$$I_k = \left\{i : \mathbf{U}_{k,[i,:]} \neq 0\right\}, \quad k = 1,2,3,$$
$$\zeta_j = (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)^\top \text{vec}(\boldsymbol{\mathcal{X}}_j^{(1)}) = \text{vec}([\![\boldsymbol{\mathcal{X}}_j^{(1)}; \mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{U}_3^\top]\!]) \in \mathbb{R}^{r_1r_2r_3}, \quad j = 1,\ldots,n_1, \tag{105}$$
$$\widetilde{\sigma}_\zeta^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \zeta_j^\top \text{vec}(\boldsymbol{\mathcal{S}})\right)^2. \tag{106}$$

Step 1 In this first step, we develop the perturbation bound for $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{W}}_k$. First, $\widetilde{\mathbf{A}}$ can be decomposed as

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{A}}} =& \frac{1}{n_1}\sum_{j=1}^{n_1} y_j^{(1)}\boldsymbol{\mathcal{X}}_j^{(1)} = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \langle \boldsymbol{\mathcal{X}}_j^{(1)}, \boldsymbol{\mathcal{A}}\rangle\right)\boldsymbol{\mathcal{X}}_j^{(1)}\\
=& \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \langle [\![\boldsymbol{\mathcal{X}}_j^{(1)}; \mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{U}_3^\top]\!], \boldsymbol{\mathcal{S}}\rangle\right)\boldsymbol{\mathcal{X}}_j^{(1)}\\
=& \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \langle [\![\boldsymbol{\mathcal{X}}_j^{(1)}; \mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{U}_3^\top]\!], \boldsymbol{\mathcal{S}}\rangle\right)[\![\boldsymbol{\mathcal{X}}_j^{(1)}; P_{\mathbf{U}_1}, P_{\mathbf{U}_2}, P_{\mathbf{U}_3}]\!]\\
&+ \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \langle [\![\boldsymbol{\mathcal{X}}_j^{(1)}; \mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{U}_3^\top]\!], \boldsymbol{\mathcal{S}}\rangle\right)P_{(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)_\perp}[\boldsymbol{\mathcal{X}}_j^{(1)}]\\
:=& \boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{R}}.
\end{aligned} \tag{107}$$

In particular, $\boldsymbol{\mathcal{H}}$ is fully determined by ζ_j and $\varepsilon_j^{(1)}$; $\boldsymbol{\mathcal{H}}$ is of Tucker rank-(p_1,p_2,p_3) and has loadings $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$. By Lemma 4,

$$\begin{aligned}
&\|\mathcal{M}_1(\boldsymbol{\mathcal{H}}) - \mathbf{A}_1\| = \left\|\mathbf{U}_1^\top\mathcal{M}_1(\boldsymbol{\mathcal{H}})(\mathbf{U}_3\otimes\mathbf{U}_2) - \mathbf{U}_1^\top\mathbf{A}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\right\|\\
=&\left\|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \langle [\![\boldsymbol{\mathcal{X}}_j^{(1)}; \mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{U}_3^\top]\!], \boldsymbol{\mathcal{S}}\rangle\right)\mathbf{U}_1^\top\mathbf{X}_{jk}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2) - \mathbf{U}_1^\top\mathbf{A}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\right\|\\
=&\left\|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\varepsilon_j^{(1)} + \left\langle\mathbf{U}_1^\top\mathbf{X}_{j1}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2), \mathbf{S}_1\right\rangle\right)\mathbf{U}_1^\top\mathbf{X}_{j1}^{(1)}(\mathbf{U}_3\otimes\mathbf{U}_2) - \mathbf{S}_1\right\|\\
\leq&\sqrt{\frac{(r_1 + r_2r_3)\widetilde{\sigma}^2\log p}{n_1}}
\end{aligned} \tag{108}$$

with probability at least $1-p^{-C}$. Similar inequalities also hold for $\|\mathcal{M}_2(\boldsymbol{\mathcal{H}})-\mathbf{A}_2\|$ and $\|\mathcal{M}_3(\boldsymbol{\mathcal{H}})-\mathbf{A}_3\|$. Provided that $\lambda_0=\min_{k=1,2,3}\sigma_{r_k}(\mathbf{A}_k)$ satisfies $\lambda_0^2\geq C\widetilde{\sigma}^2(r_1r_2+r_2r_3+r_3r_1)/n_1$, we have

$$\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{H}}))\geq\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))-\|\mathcal{M}_k(\boldsymbol{\mathcal{H}})-\mathbf{A}_k\|\geq(1-c)\lambda_k \tag{109}$$

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

Recall the definition of ζ_j and $\widetilde{\sigma}_\zeta^2$ in (105) (106). For any $j=1,\ldots,n_1$, $\varepsilon_j^{(1)}+\zeta_j^\top\text{vec}(\boldsymbol{\mathcal{S}})\sim N(0,\sigma^2+\|\boldsymbol{\mathcal{S}}\|_{\text{HS}}^2)\sim N(0,\sigma^2+\|\boldsymbol{\mathcal{A}}\|_{\text{HS}}^2)\sim N(0,\widetilde{\sigma}^2)$, which means $\widetilde{\sigma}_\zeta^2\sim\frac{\widetilde{\sigma}^2}{n_1}\chi_{n_1}^2$. By the tail bound of χ^2 distribution [72, Lemma 1],

$$\left|\widetilde{\sigma}_\zeta^2-\widetilde{\sigma}^2\right|\leq C\widetilde{\sigma}^2\left(\sqrt{\frac{\log p}{n_1}}+\frac{\log p}{n_1}\right)\leq C\widetilde{\sigma}^2\sqrt{\frac{\log p}{n_1}} \tag{110}$$

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

Since $\text{vec}(\boldsymbol{\mathcal{X}}_j^{(1)})$ has i.i.d. Gaussian entries and $(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)$ is orthogonal to $(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)_\perp$, we have that $(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)^\top\text{vec}(\boldsymbol{\mathcal{X}}_j^{(1)})$ is independent of $(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)_\perp^\top\text{vec}(\boldsymbol{\mathcal{X}}_j^{(1)})$ and $\boldsymbol{\mathcal{R}}$ (defined in (107)) is Gaussian distributed conditioning on fixed values of ζ_j and $\varepsilon_j^{(1)}$:

$$\begin{aligned}&\text{vec}(\boldsymbol{\mathcal{R}})\Big|\{\varepsilon_j^{(1)},\zeta_j\}_{j=1}^{n_1}\text{ has same distribution as }P_{(\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1)_\perp}\text{vec}(\boldsymbol{\mathcal{R}}_0),\\&\text{where }\boldsymbol{\mathcal{R}}_0\in\mathbb{R}^{p_1\times p_2\times p_3},\quad\boldsymbol{\mathcal{R}}_0\overset{iid}{\sim}N\left(0,\frac{\widetilde{\sigma}_\zeta^2}{n_1}\right).\end{aligned} \tag{111}$$

Particularly, $\boldsymbol{\mathcal{R}}_{[I_1,I_2,I_3]^c}\Big|\{\varepsilon_j^{(1)},\zeta_j\}_{j=1}^{n_1}\overset{iid}{\sim}N(0,\widetilde{\sigma}_\zeta)^2$, i.e., $\boldsymbol{\mathcal{R}}$ is i.i.d. Gaussian outside of the support of $\boldsymbol{\mathcal{A}}$.

Step 2 The rest of this proof will be conditioning on the fixed value of $\{\varepsilon_j^{(1)},\zeta_j\}_{j=1}^{n_1}$ that satisfies (108), (109), and (110). Provided (109), (110), and

$$n_1\geq\frac{C\widetilde{\sigma}^2}{\lambda_0^2}\left(s_1s_2s_3\log p+\sum_{k=1}^3(s_k^2r_k^2+r_{k+1}^2r_{k+2}^2)\right),$$

we have the following signal-noise-ratio assumption for denoising problem: $\widetilde{\mathbf{A}}=\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}}$,

$$\min_k\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{H}}))\geq\frac{C\widetilde{\sigma}_\zeta}{\sqrt{n_1}}\left((s_1s_2s_3\log p)^{1/2}+\sum_{k=1}^3(s_kr_k+r_{k+1}r_{k+1})\right).$$

By [136, Theorem 4] (with mild modifications to the proof to accommodate the fact that $\boldsymbol{\mathcal{R}}_{[I_1,I_2,I_3]}$ here is projection of i.i.d. Gaussian but not exactly i.i.d. Gaussian), the

STAT-SVD with the tuning parameter $\widehat{\sigma} = \mathrm{Med}(|\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{A}}})|/0.6744)$ (where 0.6744 is the 75% quantile of standard Gaussian) yields

$$\begin{aligned}\left\|\sin\Theta(\widetilde{\mathbf{U}}_k,\mathbf{U}_k)\right\|_F \leq & \frac{C\widetilde{\sigma}_\zeta\sqrt{(s_kr_k+s_k\log(p_k))/n_1}}{\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{H}}))} \\ \overset{(109)(110)}{\leq} & \frac{C\widetilde{\sigma}\sqrt{(s_kr_k+s_k\log(p_k))/n_1}}{\lambda_k}, \quad k\in J_s,\end{aligned} \tag{112}$$

$$\left\|\sin\Theta(\widetilde{\mathbf{U}}_k,\mathbf{U}_k)\right\|_F \leq \frac{C\widetilde{\sigma}_\zeta\sqrt{p_kr_k/n_1}}{\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{H}}))} \overset{(109)(110)}{\leq} \frac{C\widetilde{\sigma}\sqrt{p_kr_k/n_1}}{\lambda_k}, \quad k\notin J_s, \tag{113}$$

$$\begin{aligned}\text{and} \quad \left\|[\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!] - \boldsymbol{\mathcal{H}}\right\|_{\mathrm{HS}}^2 \leq & \frac{C\widetilde{\sigma}_\eta^2}{n_1}\Big(r_1r_2r_3 + \sum_{k\in J_s} s_k(r_k+\log p) + \sum_{k\notin J_s} p_kr_k\Big) \\ \overset{(104)}{\leq} & \frac{C\widetilde{\sigma}^2 m_s}{n_1}\end{aligned} \tag{114}$$

with probability at least $1-p^{-C}$, where $\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3$ are the outcomes of STAT-SVD procedure. Since the leading right singular vectors of $\mathcal{M}_k\left([\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!]\right)$ and $\mathcal{M}_k(\boldsymbol{\mathcal{A}})$ are $\widetilde{\mathbf{W}}_k$ and $\mathbf{W}_k$, respectively, we have

$$\begin{aligned}\left\|\sin\Theta(\widetilde{\mathbf{W}}_k,\mathbf{W}_k)\right\|_F = & \left\|\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{W}_k\right\|_F \leq \frac{\|\widetilde{\mathbf{W}}_{k\perp}^\top\mathbf{W}_k\mathbf{W}_k^\top\mathcal{M}_k(\boldsymbol{\mathcal{H}})^\top\|_F}{\sigma_{r_k}\left(\mathbf{W}_k^\top\mathcal{M}_k(\boldsymbol{\mathcal{H}})^\top\right)} \\ = & \frac{\|\widetilde{\mathbf{W}}_{k\perp}^\top\mathcal{M}_k(\boldsymbol{\mathcal{H}})^\top\|_F}{\sigma_{r_k}\left(\mathcal{M}_k(\boldsymbol{\mathcal{H}})\right)} \overset{\text{Lemma 7}}{\leq} \frac{\left\|\mathcal{M}_k\left([\![\widetilde{\mathbf{A}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!]\right) - \mathcal{M}_k(\mathbf{A})\right\|_F}{\sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{H}}))} \\ \overset{(109)(114)}{\leq} & C\widetilde{\sigma}\frac{\sqrt{m_s/n_1}}{\lambda_k}, \quad k=1,2,3.\end{aligned}$$

$$\begin{aligned}\left\|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\right\|_F = & \left\|\mathbf{A}_k\mathbf{W}_k\mathbf{W}_k^\top\widetilde{\mathbf{W}}_{k\perp}\right\|_F \leq \left\|\mathbf{W}_k^\top\widetilde{\mathbf{W}}_{k\perp}\right\|_F\cdot\|\mathbf{A}_k\| \\ = & \left\|\sin\Theta(\widetilde{\mathbf{W}}_k,\mathbf{W}_k)\right\|_F\cdot\|\mathbf{A}_k\| \leq C\kappa\widetilde{\sigma}\sqrt{m_s/n_1}.\end{aligned}$$

Since $\widetilde{\mathbf{U}}_k$ and $\mathbf{U}_k$ are the leading left singular values of $\mathcal{M}_k\left([\![\widetilde{\boldsymbol{\mathcal{A}}}; P_{\widetilde{\mathbf{U}}_1}, P_{\widetilde{\mathbf{U}}_2}, P_{\widetilde{\mathbf{U}}_3}]\!]\right)$ and $\mathbf{A}_k$, respectively,

$$\begin{aligned}\left\|\widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{A}_k\right\|_F = & \left\|\widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{A}_k\right\|_F \leq \left\|\widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{U}_k\right\|_F\cdot\left\|\mathbf{U}_k^\top\mathbf{A}_k\right\| = \left\|\sin\Theta(\widetilde{\mathbf{U}}_k,\mathbf{U}_k)\right\|_F\cdot\|\mathbf{A}_k\| \\ \leq & \begin{cases}\frac{C\widetilde{\sigma}\sqrt{(s_kr_k+s_k\log(p_k))/n_1}}{\lambda_k}\cdot\|\mathbf{A}_k\| \leq C\kappa\widetilde{\sigma}\sqrt{(s_kr_k+s_k\log(p_k))/n_1}, & k\in J_s; \\ \frac{C\widetilde{\sigma}\sqrt{p_kr_k/n_1}}{\lambda_k}\cdot\|\mathbf{A}_k\| \leq C\kappa\widetilde{\sigma}\sqrt{p_kr_k/n_1}, & k\notin J_s.\end{cases}\end{aligned}$$

In summary, in the previous two steps, we have shown

$$
\begin{aligned}
&\left\|\sin\Theta(\widetilde{\mathbf{U}}_k,\mathbf{U}_k)\right\|_F \le \begin{cases} \frac{C\widetilde{\sigma}\sqrt{(s_kr_k+s_k\log(p_k))/n_1}}{\lambda_k}, & k\in J_s;\\ \frac{C\widetilde{\sigma}\sqrt{p_kr_k/n_1}}{\lambda_k}, & k\notin J_s,\end{cases}\\
&\left\|\widetilde{\mathbf{U}}_{k\perp}^\top\mathbf{A}_k\right\|_F \le C\kappa\widetilde{\sigma}\sqrt{m_s/n_1},\\
&\left\|\sin\Theta(\widetilde{\mathbf{W}}_k,\mathbf{W}_k)\right\|_F \le \frac{C\widetilde{\sigma}\sqrt{m_s/n_1}}{\lambda_k},\\
&\left\|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\right\|_F \le C\kappa\widetilde{\sigma}\sqrt{m_s/n_1}, \quad \text{for} \quad k=1,2,3
\end{aligned}
\tag{115}
$$

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

Step 3 Next, we move on to analyze the second batch of samples $\{\boldsymbol{\mathcal{X}}_j^{(2)},\varepsilon_j^{(2)}\}_{j=1}^{n_2}$. We first introduce the following notations,

$$
\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2=\sigma^2+\left\|P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2^2, \quad \widehat{\sigma}_{\mathbf{E}_k}^2=\sigma^2+\left\|P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2^2.
$$

In this step, we give an upper bound for $\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2$ and $\widehat{\sigma}_{\mathbf{E}_k}^2$ given (115) holds. Note that

$$
\begin{aligned}
&\left\|P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2\\
=&\left\|\mathrm{vec}(\boldsymbol{\mathcal{A}})-P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2=\left\|\boldsymbol{\mathcal{A}}-[\![\boldsymbol{\mathcal{A}};P_{\widetilde{\mathbf{U}}_1},P_{\widetilde{\mathbf{U}}_2},P_{\widetilde{\mathbf{U}}_3}]\!]\right\|_{\mathrm{HS}}\\
=&\left\|[\![\boldsymbol{\mathcal{A}};P_{\widetilde{\mathbf{U}}_1}+P_{\widetilde{\mathbf{U}}_{1\perp}},P_{\widetilde{\mathbf{U}}_2}+P_{\widetilde{\mathbf{U}}_{2\perp}},P_{\widetilde{\mathbf{U}}_3}+P_{\widetilde{\mathbf{U}}_{3\perp}}]\!]-[\![\boldsymbol{\mathcal{A}};P_{\widetilde{\mathbf{U}}_1},P_{\widetilde{\mathbf{U}}_2},P_{\widetilde{\mathbf{U}}_3}]\!]\right\|_{\mathrm{HS}}\\
\le&\left\|\boldsymbol{\mathcal{A}};P_{\widetilde{\mathbf{U}}_{1\perp}},P_{\widetilde{\mathbf{U}}_2},P_{\widetilde{\mathbf{U}}_3}\right\|_{\mathrm{HS}}+\left\|\boldsymbol{\mathcal{A}};\mathbf{I}_{p_1},P_{\widetilde{\mathbf{U}}_{2\perp}},P_{\widetilde{\mathbf{U}}_3}\right\|_{\mathrm{HS}}+\left\|\boldsymbol{\mathcal{A}};\mathbf{I}_{p_1},\mathbf{I}_{p_2},P_{\widetilde{\mathbf{U}}_{3\perp}}\right\|_{\mathrm{HS}}\\
\le&\left\|\widetilde{\mathbf{U}}_{1\perp}^\top\mathbf{A}_1\right\|_F+\left\|\widetilde{\mathbf{U}}_{2\perp}^\top\mathbf{A}_2\right\|_F+\left\|\widetilde{\mathbf{U}}_{3\perp}^\top\mathbf{A}_3\right\|_F\\
\overset{(115)}{\le}&C\kappa\widetilde{\sigma}\sqrt{m_s/n_1},
\end{aligned}
$$

$$
\begin{aligned}
&\left\|P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2=\left\|\mathrm{vec}(\boldsymbol{\mathcal{A}})-P_{\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2\\
=&\left\|\mathbf{A}_kP_{\widetilde{\mathbf{W}}_{k\perp}}\right\|_F=\left\|\mathbf{A}_k\widetilde{\mathbf{W}}_{k\perp}\right\|_F\le C\kappa\widetilde{\sigma}\sqrt{m_s/n_1}.
\end{aligned}
$$

Therefore,

$$
\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2\le\sigma^2+\frac{Cm_s\kappa^2\widetilde{\sigma}^2}{n_1}, \quad \widehat{\sigma}_{\mathbf{E}_k}^2\le\sigma^2+\frac{Cm_s\kappa^2\widetilde{\sigma}^2}{n_1}, \quad k=1,2,3. \tag{116}
$$

Step 4 In this step, we analyze the estimation error for $\widehat{\boldsymbol{\mathcal{B}}}$ and $\widehat{\mathbf{E}}_k$ under the assumption that (115) hold (which further means (116) holds). Recall the partial linear models on importance sketching covariates (see (25) - (28); also see the proof of Theorem 3),

$$
y^{(2)}=\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{B}}})+\widetilde{\varepsilon}_{\boldsymbol{\mathcal{B}}},
$$

$$y^{(2)} = \widetilde{\mathbf{X}}_{\mathbf{E}_k}\mathrm{vec}(\widetilde{\mathbf{E}}_k) + \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}, \quad k = 1, 2, 3,$$

where the covariates, parameters, and noises of these two regressions are

$$\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n_2\times(r_1r_2r_3)}, \quad (\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})_{i\cdot} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i^{(2)} \times_1 \widetilde{\mathbf{U}}_1 \times_2 \widetilde{\mathbf{U}}_2 \times_3 \widetilde{\mathbf{U}}_3\right);$$

$$\begin{aligned}\widetilde{\mathbf{X}}_{\mathbf{E}_k} \in \mathbb{R}^{n_2\times(p_kr_k)}, \quad (\widetilde{\mathbf{X}}_{\mathbf{E}_k})_{i\cdot} =& \mathrm{vec}\left(\mathbf{X}_{ik}^{(2)}\left(\widetilde{\mathbf{U}}_{k+2}\otimes\widetilde{\mathbf{U}}_{k+1}\right)\widetilde{\mathbf{V}}_k\right)\\ =& \mathrm{vec}\left(\mathbf{X}_{ik}^{(2)}\widetilde{\mathbf{W}}_k\right), \quad k = 1, 2, 3;\end{aligned}$$

$$\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^n, \quad (\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}})_j = \left\langle \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_j^{(2)}\right); P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\rangle + \varepsilon_j^{(2)},$$

$$\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k} \in \mathbb{R}^n, \quad (\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k})_j = \left\langle \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_j^{(2)}\right), P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\rangle + \varepsilon_j^{(2)}, \quad k = 1, 2, 3;$$

$$\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{B}}}) = \mathrm{vec}([\![\boldsymbol{\mathcal{A}}; \widetilde{\mathbf{U}}_1^\top, \widetilde{\mathbf{U}}_2^\top, \widetilde{\mathbf{U}}_3^\top]\!]) = (\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\widetilde{\mathbf{U}}_1)\mathrm{vec}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{r_1r_2r_3};$$

$$\text{and} \quad \widetilde{\mathbf{E}}_k = \mathcal{M}_k\left(\boldsymbol{\mathcal{A}} \times_{k+1} \widetilde{\mathbf{U}}_{k+1}^\top \times_{k+2} \widetilde{\mathbf{U}}_{k+2}^\top\right)\widetilde{\mathbf{V}}_k = \mathbf{A}_k\mathbf{W}_k \in \mathbb{R}^{p_k\times r_k}, \quad k = 1, 2, 3.$$

These quantities satisfy the following properties.

- Based on the proof of Theorem 3, $\widetilde{\mathbf{E}}_k, k \in J_s$ are group-wise sparse,

  $$\left\|\mathrm{vec}(\widetilde{\mathbf{E}}_k)\right\|_{0,2} = \sum_{i=1}^{p_k} 1_{\left\{(\mathrm{vec}(\widetilde{\mathbf{E}}_k))_{G_i^k}\neq 0\right\}} \leq s_k,$$

  where $G_i^k = \{i + p_k, \ldots, i + p_k(r_k - 1)\}, i = 1, \ldots, p_k, k \in J_s$.
- Conditioning on fixed values of $\widetilde{\mathbf{U}}_k\widetilde{\mathbf{V}}_k, \widetilde{\mathbf{W}}_k$, the noise distribution satisfies

  $$\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}}\Big|\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k, \widetilde{\mathbf{W}}_k \overset{iid}{\sim} N\left(0, \sigma^2 + \left\|P_{(\widetilde{\mathbf{U}}_3\otimes\widetilde{\mathbf{U}}_2\otimes\mathbf{U}_1)_\perp}[\boldsymbol{\mathcal{A}}]\right\|_{\mathrm{HS}}\right) \sim N(0, \widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2);$$

  $$\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\Big|\widetilde{\mathbf{U}}_k, \widetilde{\mathbf{V}}_k, \widetilde{\mathbf{W}}_k \overset{iid}{\sim} N\left(0, \sigma^2 + \left\|P_{\left(\mathcal{R}_k(\widetilde{\mathbf{W}}_k\otimes\mathbf{I}_{p_k})\right)_\perp}[\boldsymbol{\mathcal{A}}]\right\|_{\mathrm{HS}}\right) \sim N(0, \widehat{\sigma}_{\mathbf{E}_k}^2).$$

- Note that $\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}$ is an n_2-by-$(r_1r_2r_3)$ matrix with i.i.d. Gaussian entries. Similarly to the argument in Step 5 in the proof of Theorem 4,

  $$\begin{aligned}&\left\|\left(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\right)^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}}\right\|_2^2\\ \leq&\frac{\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2\left(n_2 + 2\sqrt{n_2C\log(p)} + 2C\log(p)\right)\left(r_1r_2r_3 + 2\sqrt{Cr_1r_2r_3\log(p)} + 2C\log(p)\right)}{\left(\sqrt{n_2} - \sqrt{r_1r_2r_3} - C\log(p)\right)^4}\\ \leq&\frac{\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2}{n_2}\frac{\left(1 + 2\sqrt{\frac{C\log p}{n_2}} + 2\frac{\log p}{n_2}\right)Cm_s}{\left(1 - \sqrt{\frac{r_1r_2r_3}{n_2}} - C\sqrt{\frac{\log(p)}{n_2}}\right)^4} \leq \frac{C\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2m_s}{n_2}.\end{aligned}$$

with probability at least $1-p^{-C}$. Here, the second last inequality is due to $\sqrt{r_1r_2r_3\log(p)} \leq \frac{1}{2}\left(r_1r_2r_3+\log(p)\right) \leq m_s$ and the last inequality is due to $n_2 \geq Cm_s$. By the proof of Theorem 3,

$$\left\|\widehat{\boldsymbol{\mathcal{B}}}-\widetilde{\boldsymbol{\mathcal{B}}}\right\|_{\mathrm{HS}}^2 \overset{(65)}{=} \left\|\left(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}\right)^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top\widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}}\right\|_2^2 \leq \frac{Cm_s\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2}{n_2}. \tag{117}$$

with probability at least $1-p^{-C}$. Similarly, we can show for $k \notin J_s$, the least square estimator $\widehat{\mathbf{E}}_k$ satisfies

$$\left\|\widehat{\mathbf{E}}_k-\mathbf{E}_k\right\|_F^2 \overset{(64)}{=} \left\|\left(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top\widetilde{\mathbf{X}}_{\mathbf{E}_k}\right)^{-1}\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2 \leq \frac{Cm_s\widehat{\sigma}_{\mathbf{E}_k}^2}{n_2}. \tag{118}$$

- By Lemma 12 and $n_2 \geq Cm_s$ for large constant $C>0$, $\widetilde{\mathbf{X}}_{\mathbf{D}_k}$ satisfies group restricted isometry property with $\delta = 1/4$ with probability at least $1-\exp(-cn)$.

  Next, since $\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k} \overset{iid}{\sim} N_{n_2}\left(0,\widehat{\sigma}_{\mathbf{E}_k}^2\right)$ and $(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i)^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\Big|\|\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2 \sim N_{r_k}\left(0,\|\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_j}\|_2^2\right)$, we know

  $$\|\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2 \sim \widehat{\sigma}_{\mathbf{E}_k}^2\chi_{n_2}^2 \quad \text{and} \quad \|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i)^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2\Big|\|\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2 \sim \|\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\|_2^2\cdot\chi_{r_k}^2$$

  By the tail bound of χ^2 distribution,

  $$\begin{aligned}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i)^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2 \leq& \widehat{\sigma}_{\mathbf{E}_k}^2\left(n_2+2\sqrt{n_2C\log(p)}+2C\log(p)\right)\left(r_k+2\sqrt{r_kC\log(p)}+2C\log(p)\right)\\ \leq& Cn_2\widehat{\sigma}_{\mathbf{E}_k}^2(r_k+\log(p))\end{aligned}$$

  with probability at least $1-p^{-C}$. Since $\log(p_k) \asymp \log(p)$, we have

  $$\max_{1\leq i\leq p_k}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i)^\top\widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2 \leq Cn_2\widehat{\sigma}_{\mathbf{E}_k}^2(r_k+\log(p_k)) \tag{119}$$

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

- Similarly as the Step 5 in the proof of Theorem 4, one can show

  $$\left\|\widehat{\mathbf{E}}_k(\widetilde{\mathbf{U}}_k^\top\widehat{\mathbf{E}}_k)^{-1}\right\| \leq 1+\frac{C_1\kappa\widetilde{\sigma}}{\lambda_k}\sqrt{\frac{m_s}{n_1}}+\frac{C_2\kappa\widetilde{\sigma}}{\lambda_k}\sqrt{\frac{m_s}{n_2}} \leq 1+c, \quad k=1,2,3$$

  for constant $0<c<1/2$.

By previous arguments, we have shown the conditions of Theorem 3 hold with probability at least $1 - p^{-C}$ under the scenario of Theorem 6. Finally, Theorem 3 implies

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \\
\leq & \left(1 + \frac{C_1 \kappa \widetilde{\sigma}}{\lambda_0}\sqrt{\frac{m_s}{n_1 \wedge n_2}}\right)\left(\left\|(\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\mathcal{B}}}^\top \widetilde{\boldsymbol{\varepsilon}}_{\boldsymbol{\mathcal{B}}}\right\|_2^2 + C_2 \sum_{k \in J_s} s_k \max_{1 \leq i \leq p_k}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^i)^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}/n_2\right\|_2^2\right. \\
&\qquad \left. + \sum_{k \notin J_s}\left\|(\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\mathbf{X}}_{\mathbf{E}_k})^{-1}\widetilde{\mathbf{X}}_{\mathbf{E}_k}^\top \widetilde{\boldsymbol{\varepsilon}}_{\mathbf{E}_k}\right\|_2^2\right) \\
\overset{(a)}{\leq} & C\left(\frac{m_s(\widehat{\sigma}_{\boldsymbol{\mathcal{B}}}^2 + \widehat{\sigma}_{\mathbf{E}_k}^2)}{n_2} + C\sum_{k=1}^3 \frac{s_k(r_k + \log(p_k))\widehat{\sigma}_{\mathbf{E}_k}^2}{n_2}\right) \\
\overset{(b)}{\leq} & \frac{C_1 m_s}{n_2}\left(\sigma^2 + \frac{C_2 m_s \kappa^2 \widetilde{\sigma}^2}{n_1}\right)
\end{aligned}$$

with probability at least $1 - p^{-C}$. Here, (a) is due to (117), (118), and (119); (b) is due to (116). □

F.6 Proof of Theorem 7

This theorem gives a lower bound on the estimation error of sparse low-rank tensor regression. In order to prove the desired lower bound, we only need to prove the forthcoming (120) and (123), respectively. To prove each inequality, we first construct a series of tensor parameters $\boldsymbol{\mathcal{A}}^{(j)}$ that satisfy: (1) there are sufficient distances between $\boldsymbol{\mathcal{A}}^{(j)}$ and $\boldsymbol{\mathcal{A}}^{(l)}$ for any $j \neq l$; (2) the Kullback-Leiber divergence between the resulting observations, $\{y_i^{(j)}, \boldsymbol{\mathcal{X}}_i^{(j)}\}_{i=1}^n$ and $\{y_i^{(l)}, \boldsymbol{\mathcal{X}}_i^{(l)}\}_{i=1}^n$, are close. Finally, the lower bound is proved by an application of the generalized Fano's Lemma.

In order to prove this theorem, we only need to show

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{s},\boldsymbol{r}}} \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \max\left\{\frac{c r_1 r_2 r_3 \sigma^2}{n}, \max_{l=1,2,3} \frac{c\sigma^2\left(s_l r_l + s_l \log(ep_l/s_l)\right)}{n}\right\}.$$

1. If

$$r_1 r_2 r_3 = \max\left\{r_1 r_2 r_3, \max_{k=1,2,3}\left(s_k r_k + s_k \log(ep_k/s_k)\right)\right\},$$

we only need to prove

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{s},\boldsymbol{r}}} \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \frac{c r_1 r_2 r_3 \sigma^2}{n}, \tag{120}$$

for $r_1 r_2 r_3 \geq 9$ in order to finish the proof of this theorem. Construct $\boldsymbol{\mathcal{S}}_0$ as an r_1-by-r_2-by-r_3 tensor with i.i.d. Gaussian entries. Since $r_k \geq r_{k+1} r_{k+2}$ for $k = 1, 2, 3$, $\boldsymbol{\mathcal{S}}_0$ has Tucker rank-(r_1, r_2, r_3) with probability one. Let $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ be arbitrary fixed orthogonal matrices that satisfy

$$\mathbf{U}_k \in \mathbb{O}_{p_k,r_k}, \quad \|\mathbf{U}_k\|_{0,2} = \sum_{i=1}^{p_k} 1_{\{(\mathbf{U}_k)_{[i,:]}\neq 0\}} \leq s_k, \quad k = 1,2,3.$$

By Varshamov-Gilbert bound [87, Lemma 4.7], we can find $\boldsymbol{\mathcal{B}}^{(1)},\ldots,\boldsymbol{\mathcal{B}}^{(N)} \subseteq \{-1,1\}^{r_1\times r_2\times r_3}$ such that

$$\forall j \neq l, \quad \|\boldsymbol{\mathcal{B}}^{(j)} - \boldsymbol{\mathcal{B}}^{(l)}\|_{\mathrm{HS}}^2 = 2\sum_{i_1,i_2} |\boldsymbol{\mathcal{B}}^{(j)}_{[i_1,i_2]} - \boldsymbol{\mathcal{B}}^{(l)}_{[i_1,i_2]}| \geq 2r_1r_2r_3 \quad \text{and} \quad N \geq \exp(r_1r_2r_3/8).$$

On the other hand,

$$\|\boldsymbol{\mathcal{B}}^{(j)} - \boldsymbol{\mathcal{B}}^{(l)}\|_{\mathrm{HS}}^2 \leq 2\|\boldsymbol{\mathcal{B}}^{(j)}\|_{\mathrm{HS}}^2 + 2\|\boldsymbol{\mathcal{B}}^{(l)}\|_{\mathrm{HS}}^2 \leq 4r_1r_2r_3. \tag{121}$$

Since $r_1r_2r_3 \geq 9$, $N \geq 3$. Then we construct

$$\boldsymbol{\mathcal{A}}^{(j)} = [\![\boldsymbol{\mathcal{S}}_0 + \tau\boldsymbol{\mathcal{B}}_j; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!], \quad j = 1,\ldots,N,$$

where $\tau > 0$ is a constant to be determined a little while later. By such the configuration, $\boldsymbol{\mathcal{A}}^{(1)},\ldots,\boldsymbol{\mathcal{A}}^{(N)} \subseteq \mathcal{A}_{\boldsymbol{p},\boldsymbol{s},\boldsymbol{r}}$. Now, the KullbackLeibler divergence between the samples generated from $\boldsymbol{\mathcal{A}}^{(j)}$ and the samples generated from $\boldsymbol{\mathcal{A}}^{(l)}$ satisfy

$$\begin{aligned} &D_{KL}\left(\{\boldsymbol{\mathcal{X}}_i, y_i^{(j)}\}_{i=1}^n \middle\| \{\boldsymbol{\mathcal{X}}_i, y_i^{(l)}\}_{i=1}^n\right) \overset{\text{Lemma 13}}{=} \frac{n}{2\sigma^2}\left\|\boldsymbol{\mathcal{A}}^{(j)} - \boldsymbol{\mathcal{A}}^{(l)}\right\|_{\mathrm{HS}}^2 \\ \leq & \frac{n}{2\sigma^2}\left\|\tau\boldsymbol{\mathcal{B}}^{(j)} - \tau\boldsymbol{\mathcal{B}}^{(l)}\right\|_{\mathrm{HS}}^2 \overset{(121)}{\leq} \frac{n}{2\sigma^2}(4\tau^2 r_1r_2r_3) \end{aligned} \tag{122}$$

and

$$\forall j \neq l, \quad \left\|\boldsymbol{\mathcal{A}}^{(j)} - \boldsymbol{\mathcal{A}}^{(l)}\right\|_{\mathrm{HS}}^2 = \left\|\tau\boldsymbol{\mathcal{B}}^{(j)} - \tau\boldsymbol{\mathcal{B}}^{(l)}\right\|_{\mathrm{HS}}^2 \geq 2\tau^2 r_1r_2r_3.$$

By generalized Fano's lemma,

$$\begin{aligned} \inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{s},\boldsymbol{r}}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq & \inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\left\{\boldsymbol{\mathcal{A}}^{(1)},\ldots,\boldsymbol{\mathcal{A}}^{(N)}\right\}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \\ \geq & \tau^2 r_1r_2r_3\left(1 - \frac{2\tau^2 r_1r_2r_3 n/\sigma^2 + \log(2)}{\log(N)}\right). \end{aligned}$$

By setting $\tau^2 = \sigma^2\log(N/2.5)/(2r_1r_2r_3n)$, we have

$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{s},\boldsymbol{r}}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq c\tau^2 r_1r_2r_3 = \frac{c\sigma^2 r_1r_2r_3}{n},$$

which has shown (120) if $r_1r_2r_3 \geq 9$.

2. If
$$s_k r_k + s_k \log(ep_k/s_k) = \max\left\{ r_1r_2r_3, \max_{l=1,2,3} \left(s_l r_l + s_k \log(ep_l/s_l)\right)\right\},$$
we only need to prove
$$\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \frac{c\sigma^2\left(s_kr_k + s_k\log(ep_k/s_k)\right)}{n}, \tag{123}$$
provided that $s_kr_k + s_k\log(ep_k/s_k) \geq C$ for large constant $C > 0$. Without loss of generality we assume $k = 1$.

   To this end, we randomly generate an orthogonal matrix $\mathbf{S} \in \mathbb{O}_{r_2r_3,r_1}$ and construct $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1\times r_2\times r_3}$ such that $\mathcal{M}_1(\boldsymbol{\mathcal{S}}) = \mathbf{S}^\top$. We also construct $\mathbf{U}_2$ and $\mathbf{U}_3$ as fixed orthogonal matrices that satisfies $\|\mathbf{U}_2\|_{0,2} \leq s_2$ and $\|\mathbf{U}_3\|_{0,2} \leq s_3$. By Lemma 14, there exists $\{\mathbf{U}_1^{(k)}\}_{k=1}^N \subseteq \{1,0,-1\}^{p_1\times r_1}$ such that
$$\begin{aligned} &\|\mathbf{U}_1^{(j)}\|_{0,2} = \sum_{i=1}^{p_1} 1_{\left\{(\mathbf{U}_1^{(j)})_{[i,:]}\neq 0\right\}} \leq s_1, \quad j = 1,\ldots,N,\\ &\left\|\mathbf{U}_1^{(j)} - \mathbf{U}_1^{(l)}\right\|_{1,1} = \sum_{i,j}\left|(\mathbf{U}_1^{(j)})_{ij} - (\mathbf{U}_1^{(l)})_{ij}\right| > s_1r_1/2, \quad 1 \leq j \neq l \leq N, \end{aligned} \tag{124}$$
and $N \geq \exp\left(c(s_1r_1 + s_1\log(ep_1/s_1))\right)$. We further let
$$\boldsymbol{\mathcal{A}}^{(j)} = [\![\tau\boldsymbol{\mathcal{S}}; \mathbf{U}_1^{(j)}, \mathbf{U}_2, \mathbf{U}_3]\!], \quad j = 1,2,\ldots,N,$$
where τ is a fixed and to-be-determined value. By such the construction, for any $1 \leq j \neq l \leq N$,
$$\begin{aligned} \left\|\boldsymbol{\mathcal{A}}^{(j)} - \boldsymbol{\mathcal{A}}^{(l)}\right\|_{\mathrm{HS}}^2 =& \tau^2\left\|\mathbf{U}_1^{(j)}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\mathbf{U}_3^\top\otimes\mathbf{U}_2^\top - \mathbf{U}_1^{(l)}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\mathbf{U}_3^\top\otimes\mathbf{U}_2^\top\right\|_F^2\\ =&\tau^2\left\|\mathbf{U}_1^{(j)}\mathbf{S}^\top\mathbf{U}_3^\top\otimes\mathbf{U}_2^\top - \mathbf{U}_1^{(l)}\mathbf{S}^\top\mathbf{U}_3^\top\otimes\mathbf{U}_2^\top\right\|_F^2 = \tau^2\left\|\mathbf{U}_1^{(j)} - \mathbf{U}_1^{(l)}\right\|_F^2\\ &\quad(\text{since all entries of } \mathbf{U}_1^{(j)}, \mathbf{U}_1^{(l)} \in \{-1,0,1\})\\ \geq&\tau^2\left\|\mathbf{U}_1^{(j)} - \mathbf{U}_1^{(l)}\right\|_{1,1} > \tau^2 s_1r_1/2, \end{aligned}$$
$$\begin{aligned} \text{and}\quad & D_{KL}\left(\{\boldsymbol{\mathcal{X}}_i, y_i^{(j)}\}_{i=1}^n \middle\| \{\boldsymbol{\mathcal{X}}_i, y_i^{(l)}\}_{i=1}^n\right) = \frac{n}{2\sigma^2}\left\|\boldsymbol{\mathcal{A}}^{(j)} - \boldsymbol{\mathcal{A}}^{(l)}\right\|_{\mathrm{HS}}^2\\ =&\frac{n}{2\sigma^2}\tau^2\left\|\mathbf{U}_1^{(j)} - \mathbf{U}_1^{(l)}\right\|_F^2 \leq \frac{n\tau^2}{2\sigma^2}2\left(\|\mathbf{U}_1^{(j)}\|_2^2 + \|\mathbf{U}_1^{(l)}\|_2^2\right) \leq \frac{n\tau^2}{2\sigma^2}\cdot 4s_1r_1. \end{aligned} \tag{125}$$
By setting $\tau^2 = \sigma^2\log(N/2.5)/(2ns_1r_1)$, we have
$$\begin{aligned} &\inf_{\widehat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}}\in\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \left\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}}^2 \geq \frac{\tau^2 s_1r_1}{4}\left(1 - \frac{\frac{2n\tau^2s_1r_1}{\sigma^2} - \log(2)}{\log(N)}\right)\\ \geq&\frac{2\sigma^2\log(N/2.5)}{4ns_1r_1}\cdot\frac{s_1r_1}{4}\cdot c \geq \frac{c\sigma^2\left(s_1r_1 + s_1\log(ep_1/s_1)\right)}{n}, \end{aligned}$$
which has shown (123).

In summary of the previous two parts, we have finished the proof of this theorem. □

G Technical Lemmas

Lemma 1 (Kronecker Product, Vectorization, and Matricization)**.** *Suppose* $\mathbf{A} \in \mathbb{R}^{p_1\times p_2}$*,* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times p_2\times\ldots\times p_d}$*,* $\mathbf{B}_k \in \mathbb{R}^{p_k\times r_k}$*,* $\mathbf{B}_k' \in \mathbb{R}^{r_k\times d_k}$*,* $k=1,\ldots,d$*. Then,*

$$(\mathbf{B}_1 \otimes \cdots \otimes \mathbf{B}_d)\cdot(\mathbf{B}_1' \otimes \cdots \otimes \mathbf{B}_d') = (\mathbf{B}_1\mathbf{B}_1') \otimes \cdots \otimes (\mathbf{B}_d\mathbf{B}_d'), \tag{126}$$

$$\operatorname{vec}\left(\mathbf{B}_1^\top \mathbf{A}\mathbf{B}_2\right) = (\mathbf{B}_2^\top \otimes \mathbf{B}_1^\top)\operatorname{vec}(\mathbf{A}), \tag{127}$$

$$\operatorname{vec}\left([\![\boldsymbol{\mathcal{A}}; \mathbf{B}_1^\top,\ldots,\mathbf{B}_d^\top]\!]\right) = (\mathbf{B}_d^\top \otimes \cdots \otimes \mathbf{B}_1^\top)\operatorname{vec}(\boldsymbol{\mathcal{A}}), \tag{128}$$

$$\mathcal{M}_k\left([\![\boldsymbol{\mathcal{A}}; \mathbf{B}_1^\top,\ldots,\mathbf{B}_d^\top]\!]\right) = \mathbf{B}_k^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}})\left(\mathbf{B}_d \otimes \cdots \otimes \mathbf{B}_{k+1} \otimes \mathbf{B}_{k-1} \otimes \cdots \otimes \mathbf{B}_1\right). \tag{129}$$

Finally, for any $\mathbf{V}_k \in \mathbb{R}^{r-k\times r_k}$*,*

$$\begin{aligned}
&\operatorname{vec}\left(\mathbf{B}_k^\top \mathcal{M}_k\left(\left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{B}_1^\top,\ldots,\mathbf{B}_{k-1}^\top,\mathbf{B}_{k+1}^\top,\ldots,\mathbf{B}_d^\top\right]\!\!\right]\right)\mathbf{V}_k\right)\\
=&\mathbf{V}_k^\top\left(\mathbf{B}_d^\top \otimes \cdots \otimes \mathbf{B}_{k+1}^\top \otimes \mathbf{B}_{k-1}^\top \otimes \cdots \otimes \mathbf{B}_1^\top\right) \otimes (\mathbf{B}_k^\top)\cdot \operatorname{vec}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))
\end{aligned} \tag{130}$$

Proof of Lemma 1. See [65, 66] for the proof of (126), (128) and (129). We shall also note that (127) is the order-2 case of (128). Finally,

$$\begin{aligned}
&\operatorname{vec}\left(\mathbf{B}_k^\top \mathcal{M}_k\left(\left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{B}_1^\top,\ldots,\mathbf{B}_{k-1}^\top,\mathbf{B}_{k+1}^\top,\ldots,\mathbf{B}_d^\top\right]\!\!\right]\right)\mathbf{V}_k\right)\\
\overset{(127)}{=}&(\mathbf{V}_k^\top \otimes \mathbf{B}_k^\top)\operatorname{vec}\left(\mathcal{M}_k\left(\left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{B}_1^\top,\ldots,\mathbf{B}_{k-1}^\top,\mathbf{I}_{p_k},\mathbf{B}_{k+1}^\top,\ldots,\mathbf{B}_d^\top\right]\!\!\right]\right)\right)\\
\overset{(129)}{=}&(\mathbf{V}_k^\top \otimes \mathbf{B}_k^\top)\operatorname{vec}\left(\mathcal{M}_k(\boldsymbol{\mathcal{A}})(\mathbf{B}_d \otimes \cdots \otimes \mathbf{B}_{k+1} \otimes \mathbf{B}_{k-1} \otimes \cdots \otimes \mathbf{B}_1)\right)\\
\overset{(127)}{=}&(\mathbf{V}_k^\top \otimes \mathbf{B}_k^\top)\left(\mathbf{B}_d^\top \otimes \cdots \otimes \mathbf{B}_{k+1}^\top \otimes \mathbf{B}_{k-1}^\top \otimes \cdots \otimes \mathbf{B}_1^\top \otimes \mathbf{I}\right)\operatorname{vec}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))\\
=&\mathbf{V}_k^\top\left(\mathbf{B}_d^\top \otimes \cdots \otimes \mathbf{B}_{k+1}^\top \otimes \mathbf{B}_{k-1}^\top \otimes \cdots \otimes \mathbf{B}_1^\top\right) \otimes (\mathbf{B}_k^\top)\cdot \operatorname{vec}(\mathcal{M}_k(\boldsymbol{\mathcal{A}}))
\end{aligned}$$

□

Lemma 2. *Suppose* $\mathbf{A} \in \mathbb{R}^{p\times r}$ *and* $\mathbf{U} \in \mathbb{O}_{p,m}$*. Then,*

$$\sigma_r^2(\mathbf{A}) \geq \sigma_r^2(\mathbf{U}^\top\mathbf{A}) + \sigma_r^2(\mathbf{U}_\perp^\top\mathbf{A}), \quad \|\mathbf{A}\|^2 \leq \left\|\mathbf{U}^\top\mathbf{A}\right\|^2 + \left\|\mathbf{U}_\perp^\top\mathbf{A}\right\|^2.$$

Proof of Lemma 2. Let $\mathbf{v}$ be the right singular vector associated with the r-th singular value of $\mathbf{A}$. Then $\|\mathbf{A}\mathbf{v}\|_2 = \sigma_r(\mathbf{A})\|\mathbf{v}\|_2 = \sigma_r(\mathbf{A})$ and

$$\begin{aligned}
\sigma_r^2(\mathbf{A}) =&\|\mathbf{A}\mathbf{v}\|_2^2 = \|P_{\mathbf{U}}\mathbf{A}\mathbf{v}\|_2^2 + \|P_{\mathbf{U}_\perp}\mathbf{A}\mathbf{v}\|_2^2 = \|\mathbf{U}^\top\mathbf{A}\mathbf{v}\|_2^2 + \|\mathbf{U}_\perp^\top\mathbf{A}\mathbf{v}\|_2^2\\
\geq&\sigma_r^2(\mathbf{U}^\top\mathbf{A})\|\mathbf{v}\|_2^2 + \sigma_r^2(\mathbf{U}_\perp^\top\mathbf{A})\|\mathbf{v}\|_2^2 = \sigma_r^2(\mathbf{U}^\top\mathbf{A}) + \sigma_r^2(\mathbf{U}_\perp^\top\mathbf{A}).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\mathbf{A}\|^2 =& \max_{\mathbf{v}:\|\mathbf{v}\|_2\leq 1}\|\mathbf{A}\mathbf{v}\|_2^2 = \max_{\mathbf{v}:\|\mathbf{v}\|_2\leq 1}\left(\|P_{\mathbf{U}}\mathbf{A}\mathbf{v}\|_2^2 + \|P_{\mathbf{U}_\perp}\mathbf{A}\mathbf{v}\|_2^2\right)\\
\leq& \max_{\mathbf{v}:\|\mathbf{v}\|_2\leq 1}\|P_{\mathbf{U}}\mathbf{A}\mathbf{v}\|_2^2 + \max_{\mathbf{v}:\|\mathbf{v}\|_2\leq 1}\|P_{\mathbf{U}_\perp}\mathbf{A}\mathbf{v}\|_2^2 = \|\mathbf{U}^\top\mathbf{A}\|^2 + \|\mathbf{U}_\perp^\top\mathbf{A}\|^2.
\end{aligned}$$

□

The following lemma establish a deterministic upper bound for $\|\widehat{\mathbf{F}}\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}} - \mathbf{F}\mathbf{G}^{-1}\mathbf{H}\|$ in terms of $\|\widehat{\mathbf{F}} - \mathbf{F}\|_F, \|\widehat{\mathbf{G}} - \mathbf{G}\|_F, \|\widehat{\mathbf{H}} - \mathbf{H}\|_F$ and its more general high-order form. This result serves as a key technical lemma for the theoretical analysis of the oracle inequalities.

Lemma 3. *Suppose* $\mathbf{F}, \widehat{\mathbf{F}} \in \mathbb{R}^{p_1 \times r}, \mathbf{G}, \widehat{\mathbf{G}} \in \mathbb{R}^{r \times r}, \mathbf{H}, \widehat{\mathbf{H}} \in \mathbb{R}^{r \times p_2}$*. If* $\mathbf{G}$ *and* $\widehat{\mathbf{G}}$ *are invertible,* $\|\mathbf{F}\mathbf{G}^{-1}\| \leq \lambda_1$*,* $\|\mathbf{G}^{-1}\mathbf{H}\| \leq \lambda_2$*, and* $\|\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\| \leq \lambda_2$*, we have*

$$\left\|\widehat{\mathbf{F}}\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}} - \mathbf{F}\mathbf{G}^{-1}\mathbf{H}\right\|_F \leq \lambda_2\|\widehat{\mathbf{F}} - \mathbf{F}\|_F + \lambda_1\|\widehat{\mathbf{H}} - \mathbf{H}\|_F + \lambda_1\lambda_2\|\widehat{\mathbf{G}} - \mathbf{G}\|_F. \tag{131}$$

More generally for any $d \geq 1$*, suppose* $\widehat{\boldsymbol{\mathcal{F}}}, \boldsymbol{\mathcal{F}} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ *are order-d tensors,* $\mathbf{G}_k, \widehat{\mathbf{G}}_k \in \mathbb{R}^{r_k \times r_k}$ $\mathbf{H}_k, \widehat{\mathbf{H}}_k \in \mathbb{R}^{p_k \times r_k}$*. If* $\|\mathbf{H}_k\mathbf{G}_k^{-1}\| \leq \lambda_k$*,* $\|\widehat{\mathbf{H}}_k\widehat{\mathbf{G}}_k^{-1}\| \leq \lambda_k$*, and* $\|\mathbf{G}_k^{-1}\mathcal{M}_k(\boldsymbol{\mathcal{F}})\| \leq \pi_k$*, we have*

$$\begin{aligned}&\left\|[\![\widehat{\boldsymbol{\mathcal{F}}}; (\widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}), \ldots, (\widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1})]\!] - [\![\boldsymbol{\mathcal{F}}; (\mathbf{H}_1\mathbf{G}_1^{-1}), \ldots, (\mathbf{H}_d\mathbf{G}_d^{-1})]\!]\right\|_{\text{HS}} \\ \leq &\lambda_1 \cdots \lambda_d\|\widehat{\boldsymbol{\mathcal{F}}} - \boldsymbol{\mathcal{F}}\|_{\text{HS}} + \sum_{k=1}^{d} \pi_k\lambda_1 \cdots \lambda_d\|\widehat{\mathbf{G}} - \mathbf{G}\|_F + \sum_{k=1}^{d} \pi_k\lambda_1 \cdots \lambda_d/\lambda_k\|\widehat{\mathbf{H}}_k - \mathbf{H}_k\|_F.\end{aligned} \tag{132}$$

Proof of Lemma 3. First, it is easy to check the following identity for any non-singular matrices $\mathbf{G}$ and $\widehat{\mathbf{G}}$,

$$\widehat{\mathbf{G}}^{-1} = \mathbf{G}^{-1} - \mathbf{G}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\mathbf{G}}^{-1}.$$

Thus,

$$\begin{aligned}&\left\|\widehat{\mathbf{F}}\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}} - \mathbf{F}\mathbf{G}^{-1}\mathbf{H}\right\|_F \\ \leq &\left\|(\widehat{\mathbf{F}} - \mathbf{F})\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\right\|_F + \left\|\mathbf{F}\left(\mathbf{G}^{-1} - \mathbf{G}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\mathbf{G}}^{-1}\right)\widehat{\mathbf{H}} - \mathbf{F}\mathbf{G}^{-1}\mathbf{H}\right\|_F \\ \leq &\left\|\widehat{\mathbf{F}} - \mathbf{F}\right\|_F \cdot \left\|\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\right\| + \left\|\mathbf{F}\mathbf{G}^{-1}\widehat{\mathbf{H}} - \mathbf{F}\mathbf{G}^{-1}\mathbf{H}\right\|_F + \left\|\mathbf{F}\mathbf{G}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\right\|_F \\ \leq &\left\|\widehat{\mathbf{F}} - \mathbf{F}\right\|_F \left\|\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\right\| + \|\mathbf{F}\mathbf{G}^{-1}\| \left\|\widehat{\mathbf{H}} - \mathbf{H}\right\|_F + \|\mathbf{F}\mathbf{G}^{-1}\| \left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|_F \left\|\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{H}}\right\| \\ \leq &\lambda_2\|\widehat{\mathbf{F}} - \mathbf{F}\|_F + \lambda_1\|\widehat{\mathbf{H}} - \mathbf{H}\|_F + \lambda_1\lambda_2\|\widehat{\mathbf{G}} - \mathbf{G}\|_F.\end{aligned}$$

Then we consider the proof of (132). Define

$$\begin{aligned}\widehat{\widetilde{\mathbf{F}}}_d &= \mathcal{M}_d(\widehat{\boldsymbol{\mathcal{F}}})\left(\widehat{\mathbf{H}}_{d-1}\widehat{\mathbf{G}}_{d-1}^{-1} \otimes \cdots \otimes \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}\right)^\top, \\ \widetilde{\mathbf{F}}_d &= \mathcal{M}_d(\boldsymbol{\mathcal{F}})\left(\mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1} \otimes \cdots \otimes \mathbf{H}_1\mathbf{G}_1^{-1}\right)^\top.\end{aligned}$$

We shall note that

$$\begin{aligned}\left\|\mathbf{G}_d^{-1}\widetilde{\mathbf{F}}_d\right\| &= \left\|\mathbf{G}_d^{-1}\mathcal{M}_d(\boldsymbol{\mathcal{F}})\left(\mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1} \otimes \cdots \otimes \mathbf{H}_1\mathbf{G}_1^{-1}\right)\right\| \\ &\leq \left\|\mathbf{G}_d^{-1}\mathcal{M}_d(\boldsymbol{\mathcal{F}})\right\| \cdot \|\mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1}\| \cdots \|\mathbf{H}_1\mathbf{G}_1^{-1}\| \leq \pi_d\lambda_1 \cdots \lambda_{d-1},\end{aligned}$$

$$\left\|\mathbf{H}_d\mathbf{G}_d^{-1}\right\| \le \lambda_d, \quad \|\widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}\| \le \lambda_d.$$

By the first part of this lemma and tensor algebra,

$$\begin{aligned}
&\left\|[\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}]\!] - [\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_d\mathbf{G}_d^{-1}]\!]\right\|_{\mathrm{HS}} \\
=&\left\|\mathcal{M}_d\left([\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}]\!]\right) - \mathcal{M}_d\left([\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_d\mathbf{G}_d^{-1}]\!]\right)\right\|_F \\
\overset{\text{Lemma 1}}{=}&\left\|\widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}\widehat{\widetilde{\mathbf{F}}}_d - \mathbf{H}_d\mathbf{G}_d^{-1}\widetilde{\mathbf{F}}_d\right\|_F \\
\le&\lambda_d\|\widehat{\widetilde{\mathbf{F}}}_d - \widetilde{\mathbf{F}}_d\|_F + \lambda_1\cdots\lambda_d\pi_d\|\widehat{\mathbf{G}}_d - \mathbf{G}_d\|_F + \lambda_1\cdots\lambda_{d-1}\pi_d\|\widehat{\mathbf{H}}_d - \mathbf{H}_d\|_F.
\end{aligned} \tag{133}$$

Next, we analyze $\|\widehat{\widetilde{\mathbf{F}}}_d - \widetilde{\mathbf{F}}_d\|_F$. Define

$$\begin{aligned}
\widehat{\widetilde{\mathbf{F}}}_{d-1} &= \mathcal{M}_{d-1}(\widehat{\boldsymbol{\mathcal{F}}})\left(\mathbf{I}_{r_d}\otimes\widehat{\mathbf{H}}_{d-2}\widehat{\mathbf{G}}_{d-2}^{-1}\otimes\cdots\otimes\widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}\right)^\top, \\
\widetilde{\mathbf{F}}_{d-1} &= \mathcal{M}_{d-1}(\boldsymbol{\mathcal{F}})\left(\mathbf{I}_{r_d}\otimes\mathbf{H}_{d-2}\mathbf{G}_{d-2}^{-1}\otimes\cdots\otimes\mathbf{H}_1\mathbf{G}_1^{-1}\right)^\top.
\end{aligned}$$

Then by tensor algebra (Lemma 1),

$$\begin{aligned}
&\|\widehat{\widetilde{\mathbf{F}}}_d - \widetilde{\mathbf{F}}_d\|_F = \left\|[\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_{d-1}\widehat{\mathbf{G}}_{d-1}^{-1}, \mathbf{I}_{r_d}]\!] - [\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1}, \mathbf{I}_{r_d}]\!]\right\|_{\mathrm{HS}} \\
=&\left\|\mathcal{M}_{d-1}\left([\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_{d-1}\widehat{\mathbf{G}}_{d-1}^{-1}, \mathbf{I}_{r_d}]\!]\right) - \mathcal{M}_{d-1}\left([\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1}, \mathbf{I}_{r_d}]\!]\right)\right\|_F \\
=&\left\|\widehat{\mathbf{H}}_{d-1}\widehat{\mathbf{G}}_{d-1}^{-1}\widehat{\widetilde{\mathbf{F}}}_{d-1} - \mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1}\widetilde{\mathbf{F}}_{d-1}\right\|_F.
\end{aligned}$$

Similarly as the previous argument, one can show by the first part of this lemma that

$$\begin{aligned}
&\|\widehat{\widetilde{\mathbf{F}}}_d - \widetilde{\mathbf{F}}_d\|_F = \left\|\widehat{\mathbf{H}}_{d-1}\widehat{\mathbf{G}}_{d-1}^{-1}\widehat{\widetilde{\mathbf{F}}}_{d-1} - \mathbf{H}_{d-1}\mathbf{G}_{d-1}^{-1}\widetilde{\mathbf{F}}_{d-1}\right\|_F \\
\le&\lambda_{d-1}\|\widehat{\widetilde{\mathbf{F}}}_{d-1} - \widetilde{\mathbf{F}}_{d-1}\|_F + \lambda_1\cdots\lambda_{d-1}\pi_{d-1}\|\widehat{\mathbf{G}}_{d-1} - \mathbf{G}_{d-1}\|_F + \lambda_1\cdots\lambda_{d-2}\pi_{d-1}\|\widehat{\mathbf{H}}_{d-1} - \mathbf{H}_{d-1}\|_F.
\end{aligned}$$

Therefore, by (133) and the previous inequality,

$$\begin{aligned}
&\left\|[\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}]\!] - [\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_d\mathbf{G}_d^{-1}]\!]\right\|_{\mathrm{HS}} \\
\le&\lambda_{d-1}\lambda_d\left\|\widehat{\widetilde{\mathbf{F}}}_{d-1} - \widetilde{\mathbf{F}}_{d-1}\right\|_F + \sum_{k=d-1,d}\lambda_1\cdots\lambda_d\pi_k\|\widehat{\mathbf{G}}_k - \mathbf{G}_k\|_F + \sum_{k=d-1,d}\frac{\lambda_1\cdots\lambda_d\pi_k}{\lambda_k}\left\|\widehat{\mathbf{H}}_k - \mathbf{H}_k\right\|_F.
\end{aligned}$$

We further introduce $\widehat{\widetilde{\mathbf{F}}}_{d-2}, \widetilde{\mathbf{F}}_{d-2}, \dots, \widehat{\widetilde{\mathbf{F}}}_1, \widetilde{\mathbf{F}}_1$, repeat the previous argument for d time, and can finally obtain

$$\begin{aligned}
&\left\|[\![\widehat{\boldsymbol{\mathcal{F}}}; \widehat{\mathbf{H}}_1\widehat{\mathbf{G}}_1^{-1}, \dots, \widehat{\mathbf{H}}_d\widehat{\mathbf{G}}_d^{-1}]\!] - [\![\boldsymbol{\mathcal{F}}; \mathbf{H}_1\mathbf{G}_1^{-1}, \dots, \mathbf{H}_d\mathbf{G}_d^{-1}]\!]\right\|_{\mathrm{HS}} \\
\le&\lambda_1\cdots\lambda_d\|\widehat{\boldsymbol{\mathcal{F}}} - \boldsymbol{\mathcal{F}}\|_{\mathrm{HS}} + \sum_{k=1}^d\lambda\cdots\lambda_d\pi_k\|\widehat{\mathbf{G}}_k - \mathbf{G}_k\|_F + \sum_{k=1}^d\frac{\lambda_1\cdots\lambda_d\pi_k}{\lambda_k}\|\widehat{\mathbf{H}}_k - \mathbf{H}_k\|_F,
\end{aligned}$$

which has finished the proof of this lemma. $\square$

The following lemma characterizes the concentration of Gaussian ensemble measurements, which will be extensively used in the proof of Theorem 4.

Lemma 4 (Gaussian Ensemble Concentration Inequality for Matrices)**.** *Suppose* $\mathbf{A} \in \mathbb{R}^{a \times b}$ *is a fixed matrix,* $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{a \times b}$ *are random matrices with i.i.d. standard Gaussian entries, and* $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$*. Let* $\mathbf{E} = \frac{1}{n}\sum_{i=1}^n \left(\langle \mathbf{A}, \mathbf{X}_i \rangle + \varepsilon_i\right) \mathbf{X}_i$*. Then there exists a uniform constant* $C > 0$ *such that,*

$$\mathbb{P}\left(\|\mathbf{E} - \mathbf{A}\| \geq C\sqrt{(a+b)(\|\mathbf{A}\|_F^2 + \sigma^2)}\left(\sqrt{\frac{\log(a+b)+t}{n}} + \frac{\log(a+b)+t}{n}\right)\right) \leq \exp(-t) \tag{134}$$

Proof of Lemma 4. Denote $\mathbf{Z}_i = \left(\langle \mathbf{A}, \mathbf{X}_i \rangle + \varepsilon_i\right) \mathbf{X}_i$. It is easy to check that $\mathbb{E}\mathbf{Z}_i = \mathbf{A}$. Then,

$$\begin{aligned}
\mathbb{E}(\mathbf{Z}_i - \mathbf{A})(\mathbf{Z}_i - \mathbf{A})^\top =& \mathbb{E}\mathbf{Z}_i\mathbf{Z}_i^\top - \mathbf{A}(\mathbb{E}\mathbf{Z}_i)^\top - (\mathbb{E}\mathbf{Z}_i)\mathbf{A}^\top + \mathbf{A}\mathbf{A}^\top = \mathbb{E}\mathbf{Z}_i\mathbf{Z}_i^\top - \mathbf{A}\mathbf{A}^\top \\
=& \mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \mathbf{X}_i\mathbf{X}_i^\top + \sigma^2 \mathbb{E}\mathbf{X}_i\mathbf{X}_i^\top - \mathbf{A}\mathbf{A}^\top \\
=& \mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \mathbf{X}_i\mathbf{X}_i^\top + \sigma^2 \cdot b\mathbf{I}_a - \mathbf{A}\mathbf{A}^\top
\end{aligned}$$

Note that for any entry $(\mathbf{X}_i)_{[j,k]}$, $\mathbb{E}(\mathbf{X}_i)_{[j,k]} = 0, \mathbb{E}(\mathbf{X}_i)_{[j,k]}^2 = 1, \mathbb{E}(\mathbf{X}_i)_{[j,k]}^3 = 0, \mathbb{E}(\mathbf{X}_i)_{[j,k]}^4 = 3$. When $j \neq k$,

$$\begin{aligned}
&\left(\mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \mathbf{X}_i\mathbf{X}_i^\top\right)_{jk} = \mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \sum_{l=1}^b (\mathbf{X}_i)_{[j,l]}(\mathbf{X}_i)_{[k,l]} \\
=& \mathbb{E}\sum_{l=1}^b \left(2\mathbf{A}_{[j,l]}\mathbf{A}_{[k,l]}(\mathbf{X}_i)_{[i,l]}(\mathbf{X}_i)_{[k,l]}\right)(\mathbf{X}_i)_{[i,l]}(\mathbf{X}_i)_{[k,l]} \\
=& 2\sum_{l=1}^b \mathbf{A}_{[j,l]}\mathbf{A}_{[k,l]} = 2(\mathbf{A}\mathbf{A}^\top)_{[j,k]};
\end{aligned}$$

when $j = k$,

$$\begin{aligned}
&\left(\mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \mathbf{X}_i\mathbf{X}_i^\top\right)_{[j,j]} = \mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \sum_{l=1}^b (\mathbf{X}_i)_{[j,l]}^2 \\
=& \mathbb{E}\sum_{j'=1}^a \sum_{l'=1}^b \left(\mathbf{A}_{[j',l']}^2 (\mathbf{X}_i)_{[j',l']}^2\right) \cdot \sum_{l=1}^b (\mathbf{X}_i)_{[j,l]}^2 = \sum_{j'=1}^a \sum_{l'=1}^b (\mathbf{A}_{[j',l']}^2) \cdot b + 2\sum_{l=1}^b \mathbf{A}_{[j,l]}^2 \\
=& b\|\mathbf{A}\|_F^2 + 2(\mathbf{A}\mathbf{A}^\top)_{[j,j]}.
\end{aligned}$$

Therefore, $\mathbb{E}\langle \mathbf{A}, \mathbf{X}_i \rangle^2 \mathbf{X}_i\mathbf{X}_i^\top = 2\mathbf{A}\mathbf{A}^\top + b\|\mathbf{A}\|_F^2\mathbf{I}_a$, and

$$\left\|\mathbb{E}(\mathbf{Z}_i - \mathbf{A})(\mathbf{Z}_i - \mathbf{A})^\top\right\| = \left\|2\mathbf{A}\mathbf{A}^\top + b\|\mathbf{A}\|_F^2\mathbf{I}_a + b\sigma^2\mathbf{I}_a - \mathbf{A}\mathbf{A}^\top\right\| = \|\mathbf{A}\|^2 + b\|\mathbf{A}\|_F^2 + b\sigma^2. \tag{135}$$

Similarly, we can also show

$$\left\|\mathbb{E}(\mathbf{Z}_i-\mathbf{A})^\top(\mathbf{Z}_i-\mathbf{A})\right\| = \left\|2\mathbf{A}^\top\mathbf{A}+a\|\mathbf{A}\|_F^2\mathbf{I}_b+\sigma^2\mathbf{I}_a-\mathbf{A}^\top\mathbf{A}\right\| = \|\mathbf{A}\|^2+a\|\mathbf{A}\|_F^2+a\sigma^2. \tag{136}$$

Next, we consider the spectral norm of $\mathbf{Z}_i$ and aim to show that

$$\left\|\,\|\mathbf{Z}_i-\mathbf{A}\|\,\right\|_{\psi_1} = \inf_{u\geq 0}\left\{u : \mathbb{E}\exp\left(\frac{\|\mathbf{Z}_i-\mathbf{A}\|}{u}\right)\leq 2\right\} \leq C\left(\sqrt{a}+\sqrt{b}\right)\sqrt{\|\mathbf{A}\|_F^2+\sigma^2} \tag{137}$$

for uniform constant $C>0$. Note that $\langle\mathbf{A},\mathbf{X}_i\rangle+\varepsilon_i\sim N\left(0,\|\mathbf{A}\|_F^2+\sigma^2\right)$, $\mathbf{X}_i$ is a random matrix, by Gaussian tail bound inequality and random matrix theory (Corollary 5.35 in [122]),

$$\begin{aligned}&\mathbb{P}\left(|\langle\mathbf{A},\mathbf{X}_i\rangle+\varepsilon_i|\geq t\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}\right)\leq 2\exp(-t^2/3),\\&\mathbb{P}\left(\|\mathbf{X}_i\|\geq\sqrt{a}+\sqrt{b}+t\right)\leq\exp(-t^2/2).\end{aligned} \tag{138}$$

We set $u=C_0\left(\sqrt{a}+\sqrt{b}\right)\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}$ for large uniform constant $C_0\geq 80$. Thus, for any $x\geq 1$,

$$\begin{aligned}&\mathbb{P}\left(\|\mathbf{Z}_i-\mathbf{A}\|\geq xu\right)\leq\mathbb{P}\left(\|(\langle\mathbf{A},\mathbf{X}_i\rangle+\varepsilon_i)\mathbf{X}_i\|\geq xu-\|\mathbf{A}\|\right)\\\leq&\mathbb{P}\left(\|(\langle\mathbf{A},\mathbf{X}_i\rangle+\varepsilon_i)\mathbf{X}_i\|\geq\frac{xC_0(\sqrt{a}+\sqrt{b})}{2}\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}\right)\\\leq&\mathbb{P}\left(|\langle\mathbf{A},\mathbf{X}_i\rangle+\varepsilon_i|\geq\sqrt{\frac{xC_0}{2}\cdot\left(\|\mathbf{A}\|_F^2+\sigma^2\right)}\right)+\mathbb{P}\left(\|\mathbf{X}_i\|\geq\sqrt{\frac{xC_0}{2}}\cdot(\sqrt{a}+\sqrt{b})\right)\\\overset{(138)}{\leq}&3\exp(-C_0x/6).\end{aligned}$$

For any real valued function smooth g and non-negative random variable Y with density f_Y, the following identity holds,

$$\mathbb{E}g(Y)=\int_0^\infty g'(y)P(Y\geq y)dy.$$

Thus,

$$\begin{aligned}&\mathbb{E}\exp\left(\frac{\|\mathbf{Z}_i-\mathbf{A}\|}{u}\right)=\int_0^\infty\exp\left(x\right)\mathbb{P}\left(\frac{\|\mathbf{Z}_i-\mathbf{A}\|}{u}\geq x\right)dx\\\leq&\int_0^1\exp(u)du+\int_1^\infty\exp(x)\cdot 3\exp(-C_0x/6)dx\\\leq&\exp(1)-1+\frac{3}{C_0/6-1}\leq 2,\end{aligned}$$

which implies $\left\|\|\mathbf{Z}_i-\mathbf{A}\|\right\|_{\psi_1}\leq C_0\left(\sqrt{a}+\sqrt{b}\right)\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}$ for some uniform constant $C_0>0$.

Finally we apply the Bernstein-type matrix concentration inequality (c.f., Proposition 2 in [68] and Theorem 4 in [67]),

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i-\mathbf{A}\right\|\leq C\max\left\{\sigma_Z\sqrt{\frac{t+\log(a+b)}{n}},\right.$$
$$\left.(\sqrt{a}+\sqrt{b})\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}\log\left(\frac{C(\sqrt{a}+\sqrt{b})\sqrt{\|\mathbf{A}\|_F^2+\sigma^2}}{\sigma_Z}\right)\cdot\frac{t+\log(a+b)}{n}\right\} \tag{139}$$

with probability at least $1-\exp(-t)$. Here,

$$\begin{aligned}\sigma_Z:=&\max\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{Z}_i-\mathbf{A})(\mathbf{Z}_i-\mathbf{A})^\top\right\|^{1/2},\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{Z}_i-\mathbf{A})^\top(\mathbf{Z}_i-\mathbf{A})\right\|^{1/2}\right\}\\ =&\sqrt{\|\mathbf{A}\|^2+(a\vee b)\left(\|\mathbf{A}\|_F^2+\sigma^2\right)}.\end{aligned}$$

Noting that $\sqrt{(a\vee b)(\|\mathbf{A}\|_F^2+\sigma^2)}\leq\sigma_Z\leq\sqrt{(a\vee b+1)(\|\mathbf{A}\|_F^2+\sigma^2)}$, (139) implies (134). □

Lemma 5 (Gaussian Ensemble Concentration Inequality for Vector)**.** *Suppose* $x_1,\ldots,x_n\overset{iid}{\sim}N(0,\mathbf{I}_m)$ *are i.i.d. m-dimensional random vectors,* $\varepsilon_1,\ldots,\varepsilon_n\overset{iid}{\sim}N(0,\sigma^2)$*, and* $a\in\mathbb{R}^m$ *is a fixed vector. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\langle\mathbf{x}_i,\mathbf{a}\rangle+\varepsilon_i\right)\mathbf{x}_i-\mathbf{a}\right\|_2\leq\frac{C\sqrt{\|\mathbf{a}\|_2^2+\sigma^2}\left(\sqrt{n}+\sqrt{t}\right)\left(\sqrt{m}+\sqrt{t}\right)}{n}\right)\geq 1-5\exp(-t).$$

Proof of Lemma 5. Denote

$$\mathbf{x}_i=(x_{i1},\ldots,x_{im})^\top,\quad i=1,\ldots,n.$$

Since the distribution of Gaussian random vectors are invariant after orthogonal transformation, without loss of generality we assume $\mathbf{a}=(\theta,0,\ldots,0)$. Then

$$\frac{1}{n}\left(\sum_{i=1}^{n}\langle\mathbf{x}_i,\mathbf{a}\rangle+\varepsilon_i\right)\mathbf{x}_i-\mathbf{a}=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}(x_{i1}^2-1)\theta\\ \frac{1}{n}\sum_{i=1}^{n}x_{i1}\theta x_{i2}\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}x_{i1}\theta x_{im}\end{pmatrix}+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i:=h+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i;$$

Note that $\sum_{i=1}^n x_{i1}^2\sim\chi_n^2$, by tail bounds of χ^2 (c.f., [72, Lemma 1]),

$$\mathbb{P}\left(n-2\sqrt{nt}\leq\sum_{i=1}^{n}x_{i1}^2\right)\geq 1-\exp(-t),\quad\mathbb{P}\left(\sum_{i=1}^{n}x_{i1}^2\leq n+2\sqrt{nt}+2t\right)\geq 1-\exp(-t).$$

Conditioning on the fixed value of $\xi := \sum_{i=1}^n x_{i1}^2$, we have

$$\frac{1}{n}\sum_{i=1}^n x_{i1}\theta x_{ik}\Big|\xi \sim N\left(0, \frac{\theta^2\xi}{n^2}\right), \quad k = 2, \dots, n,$$

$$\|h\|_2^2\Big|\xi \sim \left(\frac{\xi}{n} - 1\right)^2\theta^2 + \frac{\theta^2\xi}{n^2}\chi^2_{m-1}.$$

Thus,

$$\begin{aligned}&\mathbb{P}\left(\|h\|_2^2 \geq 4\theta^2\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right)^2 + \frac{\theta^2\left(n + 2\sqrt{nt} + 2t\right)\left(m - 1 + 2\sqrt{(m-1)t} + 2t\right)}{n^2}\right)\\ \leq&\mathbb{P}\left(\xi \geq n + 2\sqrt{nt} + 2t\right) + \mathbb{P}\left(\xi \leq n - 2\sqrt{nt}\right) + \mathbb{P}\left(\frac{\theta^2\xi}{n^2}\chi^2_{m-1} \geq \frac{\theta^2\xi(m - 1 + 2\sqrt{(m-1)t} + 2t)}{n^2}\right)\\ \leq&3\exp(-t).\end{aligned}$$

Conditioning on fixed values of $\|\boldsymbol{\varepsilon}\|_2^2 = \sum_i \varepsilon_i^2$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{x}_i\right\|_2^2 \Bigg| \|\boldsymbol{\varepsilon}\|_2^2 \sim \frac{\sigma^2\|\boldsymbol{\varepsilon}\|_2^2}{n^2}\chi^2_m.$$

Additionally, $\mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2^2 \geq \sigma^2(n + 2\sqrt{nt} + 2t)\right) \leq \exp(-t)$, which means

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{x}_i\right\|_2^2 \geq \frac{\sigma^2\left(n + 2\sqrt{nt} + 2t\right)\left(m + 2\sqrt{mt} + 2t\right)}{n^2}\right)\\ \leq&\mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2^2 \geq \sigma^2(n + 2\sqrt{nt} + 2t)\right) + \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{x}_i\right\|_2^2 \Bigg| \|\boldsymbol{\varepsilon}\|_2^2 \geq \frac{\sigma^2\|\boldsymbol{\varepsilon}\|_2^2}{n^2}\left(m + 2\sqrt{mt} + 2t\right)\right)\\ \leq&2\exp(-t).\end{aligned}$$

Combining the previous two inequalities, we finally obtain

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \left(\langle \mathbf{x}_i, \mathbf{a}\rangle + \varepsilon_i\right)\mathbf{x}_i - \mathbf{a}\right\|_2 \leq \frac{C\sqrt{\theta^2 + \sigma^2}\left(\sqrt{n} + \sqrt{t}\right)\left(\sqrt{m} + \sqrt{t}\right)}{n}\right)\\ \geq&1 - 5\exp(-t).\end{aligned}$$

for constant $C > 0$. $\square$

Lemma 6. *Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^{a\times b}$ $(a \leq b)$ are i.i.d. standard Gaussian matrices, $\xi_1, \dots, \xi_n \overset{iid}{\sim} N(0, \tau^2)$, and $\mathbf{E} = \frac{1}{n}\sum_{i=1}^n \xi_i\mathbf{X}_i$. Then the largest and smallest singular values of $\mathbf{E}$ satisfies the following tail probability,*

$$\mathbb{P}\left(\sigma^2_{\max}(\mathbf{E}) \geq \tau^2\frac{n + 2\sqrt{nx} + 2x}{n^2}\left(\sqrt{a} + \sqrt{b} + \sqrt{2x}\right)^2\right) \leq 2\exp(-x),$$

$$\mathbb{P}\left(\sigma^2_{\min}(\mathbf{E}) \leq \tau^2\frac{n - 2\sqrt{nx}}{n^2}\left(\sqrt{b} - \sqrt{a} - \sqrt{2x}\right)^2\right) \leq 2\exp(-x).$$

Proof of Lemma 6. In the given setting, $\|\xi\|_2^2 = \sum_{i=1}^n \xi_i^2 \sim \tau^2\chi_n^2$, and

$$\mathbf{E} = \frac{1}{n}\sum_{i=1}^n \xi_i \mathbf{X}_i \Big| \|\xi\|_2 \overset{iid}{\sim} N\left(0, \frac{\|\xi\|_2^2}{n^2}\right).$$

By Corollary 5.35 in [122],

$$\begin{aligned} &\mathbb{P}\left(\sigma_{\max}^2(\mathbf{E}) \geq \frac{\|\xi\|_2^2}{n^2}\left(\sqrt{a}+\sqrt{b}+\sqrt{2x}\right)^2 \Big| \|\xi\|_2\right) \leq \exp(-x), \\ &\mathbb{P}\left(\sigma_{\min}^2(\mathbf{E}) \leq \frac{\|\xi\|_2^2}{n^2}\left(\sqrt{b}-\sqrt{a}-\sqrt{2x}\right)^2 \Big| \|\xi\|_2\right) \leq \exp(-x). \end{aligned} \tag{140}$$

By the tail bound of χ^2 distribution (Lemma 1 in [72]),

$$\mathbb{P}\left(\|\xi\|_2^2 \geq \tau^2\left(n + 2\sqrt{nx} + 2x\right)\right) \leq e^{-x}, \quad \mathbb{P}\left(\|\xi_2\|_2^2 \leq \tau^2\left(n - 2\sqrt{nx}\right)\right) \leq e^{-x}. \tag{141}$$

By (140) and (141), we have

$$\begin{aligned} &\mathbb{P}\left(\sigma_{\max}^2(\mathbf{E}) \geq \tau^2\frac{n+2\sqrt{nx}+2x}{n^2}\left(\sqrt{a}+\sqrt{b}+\sqrt{2x}\right)^2\right) \\ \leq &\mathbb{P}\left(\sigma_{\max}^2(\mathbf{E}) \geq \frac{\|\xi\|_2^2}{n^2}\left(\sqrt{a}+\sqrt{b}+\sqrt{2x}\right)^2 \text{ or } \|\xi\|_2^2 \geq \tau^2\left(n+2\sqrt{nx}+2x\right)\right) \\ \leq &\exp(-x)+\exp(-x) = 2\exp(-x); \end{aligned}$$

$$\begin{aligned} &\mathbb{P}\left(\sigma_{\min}^2(\mathbf{E}) \leq \tau^2\frac{n-2\sqrt{nx}}{n^2}\left(\sqrt{b}-\sqrt{a}-\sqrt{2x}\right)^2\right) \\ \leq &\mathbb{P}\left(\sigma_{\min}^2(\mathbf{E}) \leq \frac{\|\xi\|_2^2}{n^2}\left(\sqrt{b}-\sqrt{a}-\sqrt{2x}\right)^2 \text{ or } \|\xi\|_2^2 \leq \tau^2\left(n-2\sqrt{nx}\right)\right) \\ \leq &\exp(-x)+\exp(-x) = 2\exp(-x). \end{aligned}$$

□

The next lemma provides an upper bound for the projection error after perturbation, which is useful in the singular subspace perturbation analysis in the proofs of the main results.

Lemma 7 (Projection error after perturbation)**.** *Suppose* $\mathbf{A}, \mathbf{Z}$ *are two matrices of the same dimension and* $\widehat{\mathbf{U}} = \mathrm{SVD}_r(\mathbf{A}+\mathbf{Z})$*. Then,*

$$\left\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\right\| \leq \sigma_{r+1}(\mathbf{A}) + 2\|\mathbf{Z}\|, \quad \left\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\right\|_F \leq \sqrt{\sum_{k\geq r+1}\sigma_k^2(\mathbf{A})} + 2\|\mathbf{Z}\|_F.$$

In particular when $\mathrm{rank}(\mathbf{A}) \leq r$*,*

$$\left\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\right\| \leq 2\|\mathbf{Z}\|, \quad \left\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\right\|_F \leq 2\min\left\{\|\mathbf{Z}\|_F, \sqrt{r}\|\mathbf{Z}\|\right\}.$$

Proof of Lemma 7. Suppose $\mathbf{A} = \sum_k \sigma_k(\mathbf{A})\mathbf{u}_k\mathbf{v}_k^\top$ is the singular value decomposition. Then,

$$\begin{aligned}\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\| \leq & \left\|P_{\widehat{\mathbf{U}}_\perp}(\mathbf{A}+\mathbf{Z})\right\| + \|\mathbf{Z}\| = \sigma_{r+1}(\mathbf{A}+\mathbf{Z}) + \|\mathbf{Z}\| \\ = & \min_{\mathrm{rank}(\mathbf{M})\leq r} \|\mathbf{A}+\mathbf{Z}-\mathbf{M}\| + \|\mathbf{Z}\| \\ \leq & \left\|\mathbf{A}+\mathbf{Z}-\sum_{k=1}^r \sigma_k(\mathbf{A})\mathbf{u}_k\mathbf{v}_k^\top\right\| + \|\mathbf{Z}\| = \left\|\mathbf{Z}+\sum_{k\geq r+1}\sigma_k(\mathbf{A})\mathbf{u}_k\mathbf{v}_k^\top\right\| + \|\mathbf{Z}\| \\ \leq & \sigma_{r+1}(\mathbf{A}) + 2\|\mathbf{Z}\|.\end{aligned}$$

$$\begin{aligned}\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\|_F \leq & \left\|P_{\widehat{\mathbf{U}}_\perp}(\mathbf{A}+\mathbf{Z})\right\|_F + \|P_{\widehat{\mathbf{U}}_\perp}\mathbf{Z}\|_F = \sqrt{\sum_{k\geq r+1}\sigma_k^2(\mathbf{A}+\mathbf{Z})} + \|\mathbf{Z}\|_F \\ = & \min_{\mathrm{rank}(\mathbf{M})\leq r} \|\mathbf{A}+\mathbf{Z}-\mathbf{M}\|_F + \|\mathbf{Z}\|_F \\ \leq & \left\|\mathbf{A}+\mathbf{Z}-\sum_{k=1}^r \sigma_k(\mathbf{A})\mathbf{u}_k\mathbf{v}_k^\top\right\|_F + \|\mathbf{Z}\|_F \leq \sqrt{\sum_{k\geq r+1}\sigma_k^2(\mathbf{A})} + 2\|\mathbf{Z}\|_F.\end{aligned}$$

Finally, when $\mathrm{rank}(\mathbf{A}) \leq r$, $\mathrm{rank}(P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}) \leq \mathrm{rank}(\mathbf{A}) \leq r$, then

$$\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\|_F \leq \min\left\{\sqrt{\sum_{k\geq r+1}\sigma_k^2(\mathbf{A})} + 2\|\mathbf{Z}\|_F, \sqrt{r}\left\|P_{\widehat{\mathbf{U}}_\perp}\mathbf{A}\right\|\right\} \leq \min\left\{2\|\mathbf{Z}\|_F, 2\sqrt{r}\|\mathbf{Z}\|\right\}.$$

□

The Lemma 8 below provides a inequality for tensors after tensor-matrix product projections.

Lemma 8. *Suppose $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is an order-d tensor and $\widetilde{\mathbf{U}}_k \in \mathbb{O}_{p_k,r_k}$, $k = 1,\ldots,d$, are orthogonal matrices. Let $\|\cdot\|_\bullet$ be a tensor norm that satisfies sub-multiplicative inequality, i.e., $\|\boldsymbol{\mathcal{A}}\times_k \mathbf{B}\|_\bullet \leq \|\boldsymbol{\mathcal{A}}\|_\bullet \cdot \|\mathbf{B}\|$ for any tensor $\boldsymbol{\mathcal{A}}$ and matrix $\mathbf{B}$ (in particular, the tensor Hilbert-Schmitt norm satisfies this condition), we have*

$$\left\|[\![\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1},\ldots,P_{\widetilde{\mathbf{U}}_d}]\!] - \boldsymbol{\mathcal{A}}\right\|_\bullet \leq \sum_{k=1}^d \left\|\boldsymbol{\mathcal{A}}\times_k P_{\widetilde{\mathbf{U}}_{k\perp}}\right\|_\bullet.$$

Specifically,

$$\left\|[\![\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1},\ldots,P_{\widetilde{\mathbf{U}}_d}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}} = \left\|P_{(\widetilde{\mathbf{U}}_d\otimes\cdots\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2 \leq \sum_{k=1}^d \left\|\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}})\right\|_F.$$

Proof of Lemma 8. Note that

$$\begin{aligned}\boldsymbol{\mathcal{A}} = & \left[\!\!\left[\boldsymbol{\mathcal{A}}; \left(P_{\widetilde{\mathbf{U}}_1} + P_{\widetilde{\mathbf{U}}_{1\perp}}\right),\ldots,\left(P_{\widetilde{\mathbf{U}}_d} + P_{\widetilde{\mathbf{U}}_{d\perp}}\right)\right]\!\!\right] \\ = & \left[\!\!\left[\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1},\ldots,P_{\widetilde{\mathbf{U}}_d}\right]\!\!\right] + \left[\!\!\left[\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_{1\perp}},\ldots,P_{\widetilde{\mathbf{U}}_d}\right]\!\!\right] + \left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{I}_{p_1}, P_{\widetilde{\mathbf{U}}_{2\perp}},\ldots,P_{\widetilde{\mathbf{U}}_d}\right]\!\!\right] \\ & + \cdots + \left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{I}_{p_1},\mathbf{I}_{p_2},\ldots,P_{\widetilde{\mathbf{U}}_{d\perp}}\right]\!\!\right].\end{aligned}$$

Additionally, $\|P_{\widetilde{\mathbf{U}}_k}\| \leq 1, \|P_{\widetilde{\mathbf{U}}_{k\perp}}\| \leq 1$. Thus,

$$\begin{aligned}\left\|[\![\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1}, \ldots, P_{\widetilde{\mathbf{U}}_d}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\bullet} \leq & \left\|\left[\!\!\left[\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_{1\perp}}, \ldots, P_{\widetilde{\mathbf{U}}_d}\right]\!\!\right]\right\|_{\bullet} + \left\|\left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{I}_{p_1}, P_{\widetilde{\mathbf{U}}_{2\perp}}, \ldots, P_{\widetilde{\mathbf{U}}_d}\right]\!\!\right]\right\|_{\bullet} \\ & + \cdots + \left\|\left[\!\!\left[\boldsymbol{\mathcal{A}}; \mathbf{I}_{p_1}, \mathbf{I}_{p_2}, \ldots, P_{\widetilde{\mathbf{U}}_{d\perp}}\right]\!\!\right]\right\|_{\bullet} \\ \leq & \sum_{k=1}^{d} \left\|\boldsymbol{\mathcal{A}} \times_k P_{\widetilde{\mathbf{U}}_{k\perp}}\right\|_{\bullet}.\end{aligned}$$

Specifically for the Hilbert-Schmitt norm,

$$\begin{aligned}&\left\|P_{(\widetilde{\mathbf{U}}_d\otimes\cdots\otimes\widetilde{\mathbf{U}}_1)_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2 = \left\|P_{(\widetilde{\mathbf{U}}_d\otimes\cdots\otimes\widetilde{\mathbf{U}}_1)}\mathrm{vec}(\boldsymbol{\mathcal{A}}) - \mathrm{vec}(\boldsymbol{\mathcal{A}})\right\|_2 \\ \leq & \left\|[\![\boldsymbol{\mathcal{A}}; P_{\widetilde{\mathbf{U}}_1}, \ldots, P_{\widetilde{\mathbf{U}}_d}]\!] - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{HS}} \leq \sum_{k=1}^{d} \left\|\boldsymbol{\mathcal{A}} \times_k P_{\widetilde{\mathbf{U}}_{k\perp}}\right\|_{\mathrm{HS}} = \sum_{k=1}^{d} \left\|\mathcal{M}_k\left(\boldsymbol{\mathcal{A}} \times_k P_{\widetilde{\mathbf{U}}_{k\perp}}\right)\right\|_F \\ = & \left\|\widetilde{\mathbf{U}}_{k\perp}^\top \mathcal{M}_k(\boldsymbol{\mathcal{A}})\right\|_F.\end{aligned}$$

Therefore, we have finished the proof of lemma 8. □

The next Lemma 9 introduces a useful inequality for the tensor projected orthogonal to a Cross structure (i.e., $\widetilde{\mathbf{U}}$ in the statement below).

Lemma 9. *Suppose $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{S}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!]$ is a rank-(r_1, r_2, r_3) tensor. $\mathbf{U}_k \in \mathbb{O}_{p_k, r_k}$ and $\mathbf{W}_k \in \mathbb{O}_{p_{k+1}p_{k+2}, p_k}$ are the left and right singular subspaces of $\mathcal{M}_k(\boldsymbol{\mathcal{A}}) := \mathbf{A}_k$, respectively. Suppose $\widetilde{\mathbf{U}}_k \in \mathbb{O}_{p_k, r_k}$ and*

$$\widetilde{\mathbf{W}}_1 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1 \in \mathbb{O}_{p_1, r_1}, \quad \widetilde{\mathbf{W}}_2 = (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_2 \in \mathbb{O}_{p_2, r_2}, \quad \widetilde{\mathbf{W}}_3 = (\widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_3 \in \mathbb{O}_{p_3, r_3}$$

are sample estimates of $\mathbf{U}$ and $\mathbf{W}_k$, respectively. Assume $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{W}}_k$ satisfy

$$\|\sin\Theta(\widetilde{\mathbf{U}}_k, \mathbf{U}_k)\| \leq \theta_k, \quad \|\widetilde{\mathbf{U}}_{k\perp}^\top \mathbf{A}_k\|_F \leq \eta_k, \quad \|\mathbf{A}_k \widetilde{\mathbf{W}}_{k\perp}\|_F \leq \xi_k, \quad k = 1, 2, 3.$$

Let

$$\widetilde{\mathbf{U}} = \left[\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1, \quad \mathcal{R}_1(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp}), \quad \mathcal{R}_2(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp}), \quad \mathcal{R}_3(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp})\right],$$

where $\mathcal{R}_k(\cdot)$ is the row-permutation operator that matches the row indices of $\widetilde{\mathbf{W}}_k \otimes \widetilde{\mathbf{U}}_{k\perp}$ to vec$(\boldsymbol{\mathcal{A}})$ *and the actual definitions of $\mathcal{R}_k$ are provided in Section A in the supplementary materials. Recall $\widetilde{\mathbf{U}}_\perp$ is the orthogonal complement of $\mathbf{U}$. Then,*

$$\begin{aligned}\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \leq & \sum_{k=1,2,3}\left(\theta_k^2\xi_k^2 + \min\{\theta_{k+1}^2\eta_{k+2}^2, \theta_{k+2}^2\eta_{k+1}^2\}\right) \\ & + \min\{\eta_1^2\theta_2^2\theta_3^2, \theta_1^2\eta_2^2\theta_3^2, \theta_1^2\theta_2^2\eta_3^2\}.\end{aligned}$$

Proof of Lemma 9. Since

$$\widetilde{\mathbf{U}} = \left[\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1, \ \ \mathcal{R}_1(\widetilde{\mathbf{W}}_1 \otimes \widetilde{\mathbf{U}}_{1\perp}), \ \ \mathcal{R}_2(\widetilde{\mathbf{W}}_2 \otimes \widetilde{\mathbf{U}}_{2\perp}), \ \ \mathcal{R}_3(\widetilde{\mathbf{W}}_3 \otimes \widetilde{\mathbf{U}}_{3\perp})\right] \in \mathbb{O}_{p_1p_2p_3,m},$$

where $m = r_1r_2r_3 + (p_1 - r_1)r_1 + (p_2 - r_2)r_2 + (p_3 - r_3)r_3$. Denote

$$\begin{aligned}
\widetilde{\mathbf{U}}_{11} &= \mathcal{R}_1\left(\left((\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_{1\perp}\right) \otimes \widetilde{\mathbf{U}}_{1\perp}\right) \in \mathbb{O}_{p_1p_2p_3,(p_1-r_1)(r_2r_3-r_1)},\\
\widetilde{\mathbf{U}}_{12} &= \mathcal{R}_2\left(\left((\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_{2\perp}\right) \otimes \widetilde{\mathbf{U}}_{2\perp}\right) \in \mathbb{O}_{p_1p_2p_3,(p_2-r_2)(r_1r_3-r_2)}, \qquad (142)\\
\widetilde{\mathbf{U}}_{13} &= \mathcal{R}_3\left(\left((\widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_1)\widetilde{\mathbf{V}}_{3\perp}\right) \otimes \widetilde{\mathbf{U}}_{3\perp}\right) \in \mathbb{O}_{p_1p_2p_3,(p_3-r_3)(r_2r_1-r_3)},
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathbf{U}}_{21} &= \widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_{2\perp} \otimes \widetilde{\mathbf{U}}_1 \in \mathbb{O}_{p_1p_2p_3,r_1(p_2-r_2)(p_3-r_3)};\\
\widetilde{\mathbf{U}}_{22} &= \widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_2 \otimes \widetilde{\mathbf{U}}_{1\perp} \in \mathbb{O}_{p_1p_2p_3,r_2(p_1-r_1)(p_3-r_3)}; \qquad (143)\\
\widetilde{\mathbf{U}}_{23} &= \widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_{2\perp} \otimes \widetilde{\mathbf{U}}_{1\perp} \in \mathbb{O}_{p_1p_2p_3,r_3(p_1-r_1)(p_2-r_2)};
\end{aligned}$$

$$\widetilde{\mathbf{U}}_{3*} = \widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_{2\perp} \otimes \widetilde{\mathbf{U}}_{1\perp} \in \mathbb{O}_{p_1p_2p_3,(p_1-r_1)(p_2-r_2)(p_3-r_3)}. \qquad (144)$$

Then it is not hard to verify that $[\widetilde{\mathbf{U}}_{11}, \widetilde{\mathbf{U}}_{12}, \widetilde{\mathbf{U}}_{13}, \widetilde{\mathbf{U}}_{21}, \widetilde{\mathbf{U}}_{22}, \widetilde{\mathbf{U}}_{23}, \widetilde{\mathbf{U}}_{3*}]$ forms an orthogonal complement of $\widetilde{\mathbf{U}}$. Thus, we have the following decomposition,

$$\|P_{\widetilde{\mathbf{U}}_\perp}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 = \sum_{k=1,2,3} \|P_{\widetilde{\mathbf{U}}_{1k}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 + \sum_{k=1,2,3} \|P_{\widetilde{\mathbf{U}}_{2k}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2 + \|P_{\widetilde{\mathbf{U}}_{3*}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_2^2.$$

We analyze each term separately as follows.

- Note that
  $$\left[(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_1, \ \ (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_{1\perp}, \ \ (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)_\perp\right]$$
  is a square orthogonal matrix, we know
  $$\left[(\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_{1\perp}, \ \ (\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)_\perp\right]$$
  is an orthogonal complement to $\widetilde{\mathbf{W}}_1$. Given the left and right singular subspaces of $\mathbf{A}_1$ are $\mathbf{U}_1$ and $\mathbf{W}_1$, we have
  $$\begin{aligned}
  &\|P_{\widetilde{\mathbf{U}}_{11}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_F^2 \overset{(142)}{=} \left\|\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{A}_1 \left((\widetilde{\mathbf{U}}_3 \otimes \widetilde{\mathbf{U}}_2)\widetilde{\mathbf{V}}_{1\perp}\right)\right\|_F^2\\
  \leq &\left\|\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{A}_1 \widetilde{\mathbf{W}}_{1\perp}\right\|_F^2 = \left\|\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{A}_1 \widetilde{\mathbf{W}}_{1\perp}\right\|_F^2 \leq \|\widetilde{\mathbf{U}}_{1\perp}^\top \mathbf{U}_1\|^2 \cdot \|\mathbf{A}_1 \widetilde{\mathbf{W}}_{1\perp}\|_F^2\\
  \leq &\|\sin\Theta(\widetilde{\mathbf{U}}_1, \mathbf{U}_1)\|^2 \cdot \|\mathbf{A}_1 \widetilde{\mathbf{W}}_{1\perp}\|_F^2 \leq \theta_1^2 \xi_1^2.
  \end{aligned}$$
  Similar inequalities also hold for $\|P_{\widetilde{\mathbf{U}}_{12}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_F^2$ and $\|P_{\widetilde{\mathbf{U}}_{13}}\mathrm{vec}(\boldsymbol{\mathcal{A}})\|_F^2$.

- $$\begin{aligned}
&\|P_{\widetilde{\mathbf{U}}_{21}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 = \|\widetilde{\mathbf{U}}_{21}^\top\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 = \|\boldsymbol{\mathcal{A}} \times_1 \widetilde{\mathbf{U}}_1^\top \times_2 \widetilde{\mathbf{U}}_{2\perp}^\top \times_3 \widetilde{\mathbf{U}}_{3\perp}^\top\|_{\text{HS}}^2\\
=&\|\widetilde{\mathbf{U}}_{2\perp}^\top\mathbf{A}_2(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_1)\|_F^2 = \|\widetilde{\mathbf{U}}_{2\perp}^\top\mathbf{U}_2\mathbf{U}_2^\top\mathbf{A}_2(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_1)\|_F^2\\
\le&\|\widetilde{\mathbf{U}}_{2\perp}^\top\mathbf{U}_2\|^2 \cdot \|\mathbf{U}_2^\top\mathbf{A}_2(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_1)\|_F^2\\
\le&\|\sin\Theta(\widetilde{\mathbf{U}}_2, \mathbf{U}_2)\|^2 \cdot \|\mathbf{A}_2(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_1)\|_F^2 = \theta_2^2 \cdot \|\boldsymbol{\mathcal{A}} \times_1 \widetilde{\mathbf{U}}_1^\top \times_3 \widetilde{\mathbf{U}}_{3\perp}^\top\|_{\text{HS}}^2\\
=&\theta_2^2 \cdot \|\widetilde{\mathbf{U}}_{3\perp}^\top\mathbf{A}_3\|_F^2 \le \theta_2^2\eta_3^2.
\end{aligned}$$

  By symmetry, $\|P_{\widetilde{\mathbf{U}}_{21}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \le \theta_3^2\eta_2^2$. Similar inequalities also hold for $\|P_{\widetilde{\mathbf{U}}_{22}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2$ and $\|P_{\widetilde{\mathbf{U}}_{23}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2$. Therefore,

  $$\|P_{\widetilde{\mathbf{U}}_{2k}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \le \min\{\theta_{k+1}^2\eta_{k+2}^2, \theta_{k+2}^2\eta_{k+1}^2\}, \quad \text{for} \;\; k = 1, 2, 3. \tag{145}$$

- Similarly as the previous part,

  $$\begin{aligned}
\|P_{\widetilde{\mathbf{U}}_{3*}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2 \le& \left\|\boldsymbol{\mathcal{A}} \times_1 \widetilde{\mathbf{U}}_{1\perp}^\top \times_2 \widetilde{\mathbf{U}}_{2\perp}^\top \times_3 \widetilde{\mathbf{U}}_{3\perp}^\top\right\|_{\text{HS}}\\
=& \left\|\widetilde{\mathbf{U}}_{1\perp}^\top\boldsymbol{\mathcal{A}}_1\left(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_{2\perp}\right)\right\|_F\\
=& \left\|\widetilde{\mathbf{U}}_{1\perp}^\top\mathbf{A}_1(\mathbf{U}_3 \otimes \mathbf{U}_2)(\mathbf{U}_3 \otimes \mathbf{U}_2)^\top(\widetilde{\mathbf{U}}_{3\perp} \otimes \widetilde{\mathbf{U}}_{2\perp})\right\|_F\\
\le& \left\|\widetilde{\mathbf{U}}_{1\perp}^\top\mathbf{A}_1(\mathbf{U}_3 \otimes \mathbf{U}_2)\right\|_F \cdot \left\|(\mathbf{U}_3^\top\widetilde{\mathbf{U}}_{3\perp}) \otimes (\mathbf{U}_2^\top\widetilde{\mathbf{U}}_{2\perp})\right\|\\
\le& \left\|\widetilde{\mathbf{U}}_{1\perp}^\top\mathbf{A}_1\right\|_F \cdot \left\|(\mathbf{U}_3^\top\widetilde{\mathbf{U}}_{3\perp})\right\| \cdot \left\|(\mathbf{U}_2^\top\widetilde{\mathbf{U}}_{2\perp})\right\| \le \eta_1\theta_2\theta_3.
\end{aligned}$$

  Similar upper bounds of $\theta_1\eta_2\theta_3$ and $\theta_1\theta_2\eta_3$ also hold. Thus,

  $$\|P_{\widetilde{\mathbf{U}}_{3*}}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \le \min\{\eta_1^2\theta_2^2\theta_3^2, \theta_1^2\eta_2^2\theta_3^2, \theta_1^2\theta_2^2\eta_3^2\}.$$

In summary,

$$\begin{aligned}
\|P_{\mathbf{U}_\perp}\text{vec}(\boldsymbol{\mathcal{A}})\|_2^2 \le& \sum_{k=1,2,3}\left(\theta_k^2\xi_k^2 + \min\{\theta_{k+1}^2\eta_{k+2}^2, \theta_{k+2}^2\eta_{k+1}^2\}\right)\\
&+ \min\{\eta_1^2\theta_2^2\theta_3^2, \theta_1^2\eta_2^2\theta_3^2, \theta_1^2\theta_2^2\eta_3^2\}.
\end{aligned}$$

□

The following lemma discusses the Bayes risk of regular linear regression. Though it is a standard result in statistical decision theory (c.f., Exercise 5.8, p. 403 in [74]), we present the proof here for completeness of statement.

Lemma 10. *Consider the linear regression model* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. *Here,* $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$; *the parameter* $\boldsymbol{\beta}$ *is generated from a prior distribution:* $\boldsymbol{\beta} \overset{iid}{\sim} N(0, \tau^2)$. *We aim to estimate* $\boldsymbol{\beta}$ *based on* $(\mathbf{y}, \mathbf{X})$ *with the minimal* ℓ_2 *risk. Then, the Bayes estimator for* $\boldsymbol{\beta}$ *and the corresponding Bayes risk are*

$$\widehat{\boldsymbol{\beta}} = \left(\frac{\sigma^2\mathbf{I}}{\tau^2} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \quad \textit{and} \quad \mathbb{E}\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2|\mathbf{X}\right) = \text{tr}\left(\left(\frac{\mathbf{I}}{\tau^2} + \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}\right)^{-1}\right).$$

Proof of Lemma 10. When $\boldsymbol{\beta} \overset{iid}{\sim} N(0, \tau^2)$ and $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$,

$$\begin{aligned}
p(\boldsymbol{\beta}\Big|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}) \\
&\propto \exp\left(-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2\sigma^2)\right) \cdot \exp(-\boldsymbol{\beta}^\top \boldsymbol{\beta}/(2\tau^2)) \\
&\propto \exp\left(-\frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{2\tau^2} - \frac{\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}}{2\sigma^2} + \frac{\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta}}{\sigma^2}\right) \qquad (146) \\
&\propto \exp\left(-\frac{1}{2}\left\|\left(\frac{\mathbf{I}}{\tau^2} + \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2}\right)^{-1/2} \frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2} - \left(\frac{\mathbf{I}}{\tau^2} + \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2}\right)^{1/2} \boldsymbol{\beta}\right\|_2^2\right)
\end{aligned}$$

Thus, the posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}\Big|\mathbf{X}, \mathbf{y} \sim N\left(\left(\frac{\sigma^2 \mathbf{I}}{\tau^2} + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \left(\frac{\mathbf{I}}{\tau^2} + \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2}\right)^{-1}\right).$$

Then, the Bayes estimator, i.e., the posterior mean, and the corresponding Bayes risk are

$$\widehat{\boldsymbol{\beta}} = \mathbb{E}(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \left(\frac{\sigma^2 \mathbf{I}}{\tau^2} + \mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}^\top \mathbf{y}, \quad \mathbb{E}((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2|\mathbf{X}, \mathbf{y}) = \operatorname{tr}\left(\left(\frac{\mathbf{I}}{\tau^2} + \frac{\mathbf{X}\mathbf{X}^\top}{\sigma^2}\right)^{-1}\right),$$

respectively. Thus, we have finished the proof of this lemma. □

The following lemma provides a deterministic bound for the group Lasso estimator under group restricted isometry property.

Lemma 11. *Suppose* $\mathbf{X} \in \mathbb{R}^{n \times pr}$, $\{G_1, \ldots, G_p\}$ *is a partition of* $\{1, \ldots, pr\}$ *and* $|G_1| = \cdots = |G_p|$. *Assume* $\mathbf{X}$ *satisfies group restricted isometry condition, such that*

$$(1 - \delta)n\|\boldsymbol{\beta}\|_2^2 \le \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \le (1 + \delta)n\|\boldsymbol{\beta}\|_2^2, \quad \forall \boldsymbol{\beta} \text{ such that } \sum_{i=1}^{m} 1_{\{\boldsymbol{\beta}_{G_i} \neq 0\}} \le 2s.$$

Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ *and* $\sum_{i=1}^{p} 1_{\{\boldsymbol{\beta}_{G_i} \neq 0\}} \le s$. *Consider the following group Lasso estimator*

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{pr}}{\arg\min} \left\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 + \eta \sum_{i=1}^{p} \|\boldsymbol{\gamma}_{G_i}\|_2\right\}. \qquad (147)$$

For $\eta \ge 3\max_{1 \le j \le p} \|(\mathbf{X}_{[:, G_j]})^\top \varepsilon\|_2$ *and* $\delta < 2/7$, *the optimal solution of* (147) *yields*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \frac{4\eta\sqrt{s/3}}{n(1 - 7\delta/2)}. \qquad (148)$$

Proof of Lemma 11. For convenience, define the $(2, \infty)$- and $(2, 1)$-norms of any vector $v \in \mathbb{R}^{pr}$ as

$$\|\mathbf{v}\|_{2,\infty} = \max_{j=1,\ldots,p} \|\mathbf{v}_{G_j}\|_2 \quad \text{and} \quad \|\mathbf{v}\|_{2,1} = \sum_{j=1}^{p} \|\mathbf{v}_{G_j}\|_2.$$

Then, $\|\cdot\|_{2,\infty}$ and $\|\cdot\|_{2,1}$ satisfies $\|\mathbf{v}\|_{2,\infty}\cdot\|w\|_{2,1}\geq\langle\mathbf{v},w\rangle$. We also define $J=\{j:\boldsymbol{\beta}_{G_j}\neq 0\}$ as the group support of $\boldsymbol{\beta}$, then $|J|\leq s$ based on the assumption. Suppose $h=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\in\mathbb{R}^{pr}$. By definition,

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2+\eta\|\widehat{\boldsymbol{\beta}}\|_{2,1}\leq\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\eta\|\boldsymbol{\beta}\|_{2,1}.$$

Noting that

$$\begin{aligned}&\frac{1}{2}\left(\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2-\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2\right)=\frac{1}{2}\left(\|\varepsilon-\mathbf{X}h\|_2^2-\|\boldsymbol{\varepsilon}\|_2^2\right)\\ =&-\frac{1}{2}(2\varepsilon-\mathbf{X}h)^\top(\mathbf{X}h)\geq-\varepsilon^\top\mathbf{X}h\geq-\|\mathbf{X}^\top\varepsilon\|_{2,\infty}\cdot\|h\|_{2,1}\\ =&-\|\mathbf{X}^\top\varepsilon\|_{2,\infty}(\|h_J\|_{2,1}+\|h_{J^c}\|_{2,1}),\end{aligned}$$

$$\eta\left(\|\boldsymbol{\beta}\|_{2,1}-\|\widehat{\boldsymbol{\beta}}\|_{2,1}\right)=\eta\left(\|\boldsymbol{\beta}_J\|_{2,1}-\|\widehat{\boldsymbol{\beta}}_J\|_{2,1}-\|\widehat{\boldsymbol{\beta}}_{J^c}\|_{2,1}\right)\leq\eta\left(\|h_J\|_{2,1}-\|h_{J^c}\|_{2,1}\right),$$

we have

$$\begin{aligned}&-\|\mathbf{X}^\top\varepsilon\|_{2,\infty}(\|h_J\|_{2,1}+\|h_{J^c}\|_{2,1})\leq\eta(\|h_J\|_{2,1}-\|h_{J^c}\|_{2,1}),\\ \Rightarrow\quad&\|h_{J^c}\|_{2,1}\leq\frac{\eta+\|\mathbf{X}^\top\varepsilon\|_{2,\infty}}{\eta-\|\mathbf{X}^\top\varepsilon\|_{2,\infty}}\|h_J\|_{2,1}.\end{aligned}$$

Given $\eta\geq 3\|\mathbf{X}^\top\varepsilon\|_{2,\infty}$, we have

$$\|h_{J^c}\|_{2,1}\leq 2\|h_J\|_{2,1}.\tag{149}$$

Now we can sort all groups of h by their ℓ_2 norm and suppose $\|h_{G_{i_1}}\|_2\geq\cdots\geq\|h_{G_{i_p}}\|_2$, where $\{i_1,\ldots,i_p\}$ as a permutation of $\{1,\ldots,p\}$. Let

$$h_{\max(s)}\in\mathbb{R}^{pr},\quad(h_{\max(s)})_j=\begin{cases}h_j, & j\in G_{i_1}\cup\cdots\cup G_{i_s};\\ 0, & \text{otherwise},\end{cases}$$

Then $h_{\max(s)}$ is the vector h with all but the s largest groups in ℓ_2 norm set to zero. We also denote $h_{-\max(s)}=h-h_{\max(s)}$. Then (149) implies

$$\|h_{-\max(s)}\|_{2,1}\leq\|h_{J^c}\|_{2,1}\leq 2\|h_J\|_{2,1}\leq 2\|h_{\max(s)}\|_{2,1}.\tag{150}$$

Let $\mathbf{v}\in\mathbb{R}^p$ with $\mathbf{v}_i=\|h_{G_i}\|_2, 1\leq i\leq p$ be the ℓ_2 norms of each group of h. We can similarly define $\mathbf{v}_{\max(s)}$ as the vector $\mathbf{v}$ with all but the s largest entries set to zero, and $\mathbf{v}_{-\max(s)}=\mathbf{v}-\mathbf{v}_{\max(s)}$. Then, $(\mathbf{v}_{\max(s)})_i=\|(h_{\max(s)})_{G_i}\|_2$ and $(\mathbf{v}_{-\max(s)})_i=\|(h_{-\max(s)})_{G_i}\|_2$. Let

$$\alpha=\max\{\|h_{-\max(s)}\|_{2,\infty},\|h_{-\max(s)}\|_{2,1}/s\}=\max\{\|\mathbf{v}_{-\max(s)}\|_\infty,\|\mathbf{v}_{-\max(s)}\|_1/s\}.$$

By the polytope representation lemma (Lemma 1 in [17]) with α, one can find a finite series of vectors $\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(N)} \in \mathbb{R}^p$ and weights $\pi_1, \ldots, \pi_N$ such that

$$\mathrm{supp}(\mathbf{v}^{(j)}) \subseteq \mathrm{supp}(\mathbf{v}_{-\max(s)}), \quad \|\mathbf{v}^{(j)}\|_0 \leq s, \quad \|\mathbf{v}^{(j)}\|_\infty \leq \alpha, \quad \|\mathbf{v}^{(j)}\|_1 = \|\mathbf{v}_{-\max(s)}\|_1,$$

$$\mathbf{v}_{-\max(s)} = \sum_{j=1}^N \pi_j \mathbf{v}^{(j)}, \quad 0 \leq \pi_j \leq 1, \quad \text{and} \quad \sum_{j=1}^N \pi_j = 1.$$

Now we construct

$$h^{(j)} \in \mathbb{R}^{pr}, \quad \text{where} \quad (h^{(j)})_{G_i} = \frac{(h_{-\max(s)})_{G_i}}{\|(h_{-\max(s)})_{G_i}\|_2} \cdot \mathbf{v}_i^{(j)}, \quad i = 1, \ldots, p; j = 1, \ldots, N. \tag{151}$$

Then $\{h^{(j)}\}_{j=1}^N$ satisfy

$$\begin{aligned} &\mathrm{supp}(h^{(j)}) \subseteq \mathrm{supp}(h_{-\max(s)}), \quad \sum_{i=1}^p 1_{\{(h^{(j)})_{G_i} \neq 0\}} \leq s, \quad \|h^{(j)}\|_{2,\infty} \leq \alpha, \\ &\|h^{(j)}\|_{2,1} = \|h_{-\max(s)}\|_{2,1}, \quad h_{-\max(s)} = \sum_{j=1}^N \pi_j h^{(j)}, \quad 0 \leq \pi_j \leq 1, \quad \sum_{j=1}^N \pi_j = 1. \end{aligned} \tag{152}$$

Therefore, $h_{\max(s)}$ and $h^{(j)}$ have distinct supports, $\sum_{i=1}^m 1_{(h_{\max(s)}+h^{(j)})_{G_i} \neq 0} \leq 2s$, $\|h_{\max(s)} + h^{(j)}\|_2^2 = \|h_{\max(s)}\|_2^2 + \|h^{(j)}\|_2^2$, and

$$\begin{aligned} \|h^{(j)}\|_2^2 \leq & \|h^{(j)}\|_{2,1} \cdot \|h^{(j)}\|_{2,\infty} \overset{(152)}{\leq} \|h_{-\max(s)}\|_{2,1} \cdot \alpha \\ \overset{(150)}{\leq} & 2\|h_{\max(s)}\|_{2,1} \cdot \max\left\{\|h_{-\max(s)}\|_{2,\infty}, \|h_{-\max(s)}\|_{2,1}/s\right\} \\ \leq & 2\|h_{\max(s)}\|_{2,1} \cdot \max\left\{\min_{j:\|h_{G_j}\|_2 \neq 0} \|h_{G_j}\|_2, 2\|h_{\max(s)}\|_{2,1}/s\right\} \\ \leq & 4\|h_{\max(s)}\|_{2,1}^2/s \leq 4\|h_{\max(s)}\|_2^2. \end{aligned}$$

Thus,

$$\begin{aligned} & \left|\langle \mathbf{X}h_{\max(s)}, \mathbf{X}h_{-\max(s)}\rangle\right| \leq \sum_{j=1}^N \pi_j \left|\langle \mathbf{X}h_{\max(s)}, \mathbf{X}h^{(j)}\rangle\right| \\ = & \sum_{j=1}^N \frac{\pi_j}{4} \left|\|\mathbf{X}h_{\max(s)} + \mathbf{X}h^{(j)}\|_2^2 - \|\mathbf{X}h_{\max(s)} - \mathbf{X}h^{(j)}\|_2^2\right| \\ \leq & \sum_{j=1}^N \frac{\pi_j}{4} \left(n(1+\delta)(\|h_{\max(s)}\|_2^2 + \|h^{(j)}\|_2^2) - n(1-\delta)(\|h_{\max(s)}\|_2^2 + \|h^{(j)}\|_2^2)\right) \\ \leq & \frac{\delta n}{2} \left(\|h_{\max(s)}\|_2^2 + 4\|h_{\max(s)}\|_2^2\right) = \frac{5\delta n}{2} \|h_{\max(s)}\|_2^2, \end{aligned}$$

which means

$$\begin{aligned}\langle \mathbf{X}h_{\max(s)}, \mathbf{X}h\rangle &= \|\mathbf{X}h_{\max(s)}\|_2^2 + \langle \mathbf{X}h_{\max(s)}, \mathbf{X}h_{-\max(s)}\rangle \\ &\geq n(1-\delta)\|h_{\max(s)}\|_2^2 - \frac{5\delta n}{2}\|h_{\max(s)}\|_2^2 = n(1-7\delta/2)\|h_{\max(s)}\|_2^2.\end{aligned} \tag{153}$$

Next, by the KKT condition of $\widehat{\boldsymbol{\beta}}$ being the optimizer of (147),

$$\|\mathbf{X}^\top(y - \mathbf{X}\widehat{\boldsymbol{\beta}})\|_{2,\infty} \leq \eta.$$

In addition, $\|\mathbf{X}^\top(y - \mathbf{X}\boldsymbol{\beta})\|_{2,\infty} = \|\mathbf{X}^\top\varepsilon\|_{2,\infty} \leq \eta/3$, which means

$$\begin{aligned}\langle \mathbf{X}h_{\max(s)}, \mathbf{X}h\rangle &= h_{\max(s)}^\top \mathbf{X}^\top\mathbf{X}h \leq \|h_{\max(s)}\|_{2,1} \cdot \|\mathbf{X}^\top\mathbf{X}h\|_{2,\infty} \\ &\leq \|h_{\max(s)}\|_{2,1} \cdot \Big(\|\mathbf{X}^\top(y - \mathbf{X}\widehat{\boldsymbol{\beta}})\|_{2,\infty} + \|\mathbf{X}^\top(y - \mathbf{X}\boldsymbol{\beta})\|_{2,\infty}\Big) \\ &\leq 4\eta/3 \cdot \|h_{\max(s)}\|_{2,1} \leq 4\eta/3 \cdot \sqrt{s}\|h_{\max(s)}\|_2.\end{aligned} \tag{154}$$

Combining the above inequality with (153), one has

$$\frac{4\eta}{3}\sqrt{s}\|h_{\max(s)}\|_2 \geq n(1-7\delta/2)\|h_{\max(s)}\|_2^2,$$

namely

$$\|h_{\max(s)}\|_2 \leq \frac{\frac{4}{3}\eta\sqrt{s}}{n(1-7\delta/2)}.$$

Finally,

$$\begin{aligned}\|h_{-\max(s)}\|_2^2 \leq &\|h_{-\max(s)}\|_{2,1} \cdot \|h_{-\max(s)}\|_{2,\infty} \\ \leq &2\|h_{\max(s)}\|_{2,1} \cdot \min_{j:(h_{\max(s)})_{G_j}\neq 0} \|(h_{\max(s)})_{G_j}\|_2 \\ \leq &2\|h_{\max(s)}\|_2^2.\end{aligned}$$

Therefore,

$$\|h\|_2 = \sqrt{\|h_{-\max(s)}\|_2^2 + \|h_{\max(s)}\|_2^2} \leq \sqrt{3}\|h_{\max(s)}\|_2 \leq \frac{4\eta\sqrt{s/3}}{n(1-7\delta/2)},$$

which has finished the proof of this lemma. □

The next Lemma 12 shows that the Gaussian Ensemble satisfies group restricted isometry property with high probability.

Lemma 12. *Suppose* $\mathbf{X} \in \mathbb{R}^{n\times(pr)}$, $G_1, \ldots, G_p$ *is a partition of* $\{1, \ldots pr\}$ *and* $|G_1| = \cdots |G_p| = r$. *If* $\mathbf{X} \overset{iid}{\sim} N(0,1)$ *and* $n \geq C(sr/\delta + s\log(ep/s))$ *for large constant* $C > 0$, $\mathbf{X}$ *satisfies the following group restricted isometry (GRIP)*

$$n(1-\delta)\|\boldsymbol{\beta}\|_2^2 \leq \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq n(1+\delta)\|\boldsymbol{\beta}\|_2^2, \quad \forall\boldsymbol{\beta} \text{ such that } \sum_{i=1}^p 1_{\{\boldsymbol{\beta}_{\mathbf{G}_i}\neq 0\}} \leq s \tag{155}$$

with probability at least $1 - \exp(-cn)$.

Proof of Lemma 12. First, the statement (155) is equivalently to

$$\begin{aligned}&\forall \text{ distinct } i_1,\ldots,i_s \subseteq \{1,\ldots,p\},\\ &n(1-\delta) \leq \sigma_{\min}^2(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq \sigma_{\max}^2(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq n(1+\delta).\end{aligned} \tag{156}$$

Since $\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}$ is an n-by-sr matrix with i.i.d. Gaussian entries, by random matrix theory (c.f., [122, Corollary 5.35]),

$$\begin{aligned}&\mathbb{P}\left(\sqrt{n}-\sqrt{sr}-x \leq \sigma_{\min}(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq \sigma_{\max}(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq \sqrt{n}+\sqrt{sr}+x\right)\\ \geq& 1-2\exp(-x^2/2),\end{aligned}$$

which means

$$\begin{aligned}&\mathbb{P}\left((156) \text{ does not hold}\right)\\ \leq& \sum_{\substack{\text{distinct } i_1,\ldots,i_s\\ \subseteq\{1,\ldots,p\}}} \mathbb{P}\left(\left\{n(1-\delta) \leq \sigma_{\min}^2(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq \sigma_{\max}^2(\mathbf{X}_{[:,G_{i_1}\cup\cdots\cup G_{i_s}]}) \leq n(1+\delta)\right\}^c\right)\\ \leq& 2\binom{p}{s}\exp\left(-\left(\sqrt{n}-\sqrt{n(1-\delta)}-\sqrt{sr}\right)_+^2 \wedge \left(\sqrt{n(1+\delta)}-\sqrt{n}-\sqrt{sr}\right)_+^2\right),\end{aligned}$$

Provided that $n \geq C(sr/\delta + s\log(ep/s))$ for large constant $C>0$, we have

$$\left(\sqrt{n}-\sqrt{n(1-\delta)}-\sqrt{sr}\right)_+^2 \wedge \left(\sqrt{n(1+\delta)}-\sqrt{n}-\sqrt{sr}\right)_+^2 \geq (1-c)n,$$

$$(1-c)n \geq (1-c)Cs\log(ep/s) \geq (1-c)C\log\left(\binom{p}{s}\right).$$

Therefore, we have

$$\mathbb{P}\left((156) \text{ does not hold}\right) \leq \exp\left(\log\left(2\binom{p}{s}\right) - (1-c)n\right) \leq \exp(-cn)$$

and have finished the proof of this lemma. □

The next lemma gives the KullbackLeibler divergence between two regression models with random designs, which will be used in the lower bound argument in this paper.

Lemma 13. *Consider two linear regression models $\mathbf{y}^{(1)} = \mathbf{X}\boldsymbol{\beta}^{(1)} + \varepsilon$ and $y^{(2)} = \mathbf{X}\boldsymbol{\beta}^{(2)} + \varepsilon$. Here, $\mathbf{y}^{(1)}, y^{(2)} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \in \mathbb{R}^p$, and $\varepsilon \in \mathbb{R}^n$. Assume $\mathbf{X} \overset{iid}{\sim} N(0,1)$, $\varepsilon \overset{iid}{\sim} N(0,\sigma^2)$, and $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}$ are fixed. Then,*

$$D_{KL}\left(\{\mathbf{X},\mathbf{y}^{(1)}\}\middle\|\{\mathbf{X},y^{(2)}\}\right) = \frac{n}{2\sigma^2}\left\|\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}\right\|_2^2. \tag{157}$$

Proof of Lemma 13. Denote the j-th row vector of $\mathbf{X}$ as x_j, i.e., $\mathbf{X} = [x_1^\top \cdots x_n^\top]^\top$. Then, $(x_1^\top, y_1^{(1)\top}), \ldots, (x_n^\top, y_n^{(1)\top})$ are i.i.d. distributed vectors, $y_j^{(1)} = x_j^\top\boldsymbol{\beta}^{(1)} + \varepsilon_j$, and

$$\left(x_j^\top, y_j^{(1)}\right) \sim N\left(0, \Sigma_1\right), \quad \Sigma_1 = \begin{bmatrix}\mathbf{I}_p & \boldsymbol{\beta}^{(1)}\\ \boldsymbol{\beta}^{(1)\top} & \|\boldsymbol{\beta}^{(1)}\|_2^2+\sigma^2\end{bmatrix}.$$

Similarly,

$$\left(x_j^\top, y_j^{(2)}\right) \sim N(0, \Sigma_2), \quad \Sigma_2 = \begin{bmatrix} \mathbf{I}_p & \boldsymbol{\beta}^{(2)} \\ \boldsymbol{\beta}^{(2)\top} & \|\boldsymbol{\beta}^{(2)}\|_2^2 + \sigma^2 \end{bmatrix}.$$

Additionally,

$$\det(\Sigma_i) = \det\left(\begin{bmatrix} \mathbf{I}_p & 0 \\ -\boldsymbol{\beta}^{(i)\top} & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{I}_p & \boldsymbol{\beta}^{(i)} \\ \boldsymbol{\beta}^{(i)\top} & \|\boldsymbol{\beta}^{(i)}\|_2^2 + \sigma^2 \end{bmatrix}\right) = \det\left(\begin{bmatrix} \mathbf{I}_p & \boldsymbol{\beta}^{(i)} \\ 0 & \sigma^2 \end{bmatrix}\right) = \sigma^2, \quad i = 1, 2,$$

$$\Sigma_i^{-1} = \begin{bmatrix} \mathbf{I}_p + \boldsymbol{\beta}^{(i)}\boldsymbol{\beta}^{(i)\top}\sigma^{-2} & -\boldsymbol{\beta}^{(i)}\sigma^{-2} \\ -\boldsymbol{\beta}^{(i)\top}\sigma^{-2} & \sigma^{-2} \end{bmatrix}, \quad i = 1, 2.$$

By the formula for multivariate normal distribution KL-divergence,

$$\begin{aligned}
& D_{KL}\left(\left\{x_j^\top, y_j^{(1)}\right\} \middle\| \left\{x_j^\top, y_j^{(2)}\right\}\right) \\
=& \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) - (p+1) + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right) \\
=& \frac{\sigma^{-2}}{2}\left(\operatorname{tr}\left(\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\top} - \boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(1)^\top} - \boldsymbol{\beta}^{(1)\top}\boldsymbol{\beta}^{(2)}\right) + \|\boldsymbol{\beta}^{(1)}\|_2^2\right) \\
=& \frac{1}{2\sigma^2}\left\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\right\|_2^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
D_{KL}\left(\left\{x_j^\top, y_j^{(1)}\right\}_{j=1}^n \middle\| \left\{x_j^\top, y_j^{(2)}\right\}_{j=1}^n\right) =& n D_{KL}\left(\left\{x_j^\top, y_j^{(1)}\right\} \middle\| \left\{x_j^\top, y_j^{(2)}\right\}\right) \\
=& \frac{n}{2\sigma^2}\left\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\right\|_2^2.
\end{aligned}$$

□

The next lemma can be seen as a sparse version of Varshamov-Gilbert bound [87, Lemma 4.7]. This result is crucial in the proof of the lower bound argument in sparse tensor regression (Theorem 7).

Lemma 14. *There exists a series of matrices* $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)} \in \{1, 0, -1\}^{p \times r}$*, such that*

$$\|\mathbf{A}^{(k)}\|_{0,2} := \sum_{i=1}^p 1_{\left\{\mathbf{A}_{[i,:]}^{(k)} \neq 0\right\}} \leq s, \quad \|\mathbf{A}^{(k)} - \mathbf{A}^{(l)}\|_{1,1} = \sum_{i=1}^p \sum_{j=1}^r \left|\mathbf{A}_{[i,j]}^{(k)} - \mathbf{A}_{[i,j]}^{(l)}\right| > sr/2 \quad (158)$$

for all k, l*, and* $N \geq \exp\left(c(sr + s\log(ep/s))\right)$ *for some uniform constant* $c > 0$.

Proof of Lemma 14 First, if $p/s \leq C$ for some constant $C > 0$, the lemma directly follows from the Varshamov-Gilbert bound by restricting on the top $s \times r$ submatrices of $\mathbf{A}_1, \ldots, \mathbf{A}_N$. Thus, without loss of generality, we assume $p \geq 10s$ throughout the rest of the proof.

Next for $k = 1, \ldots, N$, we randomly draw s elements from $\{1, \ldots, p\}$ without replacement, form $\Omega^{(k)}$ as a random subset of $\{1, \ldots, p\}$, and generate

$$\mathbf{A}^{(k)} \in \mathbb{R}^{p \times r}, \quad \left(\mathbf{A}^{(k)}\right)_{ij} \begin{cases} \sim \text{Rademacher}, & i \in \Omega^{(k)}; \\ = 0, & i \notin \Omega^{(k)}, \end{cases}$$

for $k = 1, 2, \ldots, N$. Here, $A \sim$ Rademacher if A is equally distributed on -1 and 1. By such the construction,

$$\|\mathbf{A}^{(k)}\|_{0,2} = \sum_{i=1}^{p} 1_{\left\{\mathbf{A}^{(k)}_{[i,:]} \neq 0\right\}} \leq s.$$

For any $k \neq l$,

$$\begin{aligned} \left\|\mathbf{A}^{(k)} - \mathbf{A}^{(l)}\right\|_{1,1} \sim & r|\Omega^{(k)} \backslash \Omega^{(l)}| + r|\Omega^{(l)} \backslash \Omega^{(k)}| + 2 \cdot \text{Bin}\left(r\left|\Omega^{(k)} \cap \Omega^{(l)}\right|, 1/2\right) \\ = & 2sr - 2r|\Omega^{(l)} \cap \Omega^{(k)}| - 2 \cdot \text{Bin}\left(r|\Omega^{(l)} \cap \Omega^{(k)}|, 1/2\right) \\ \sim & 2sr - 2 \cdot \text{Bin}\left(r|\Omega^{(l)} \cap \Omega^{(k)}|, 1/2\right). \end{aligned} \tag{159}$$

Here, we used the fact that $|\Omega^{(k)} \backslash \Omega^{(l)}| = |\Omega^{(k)}| - |\Omega^{(k)} \cap \Omega^{(l)}| = s - |\Omega^{(k)} \cap \Omega^{(l)}|$. Moreover, $|\Omega^{(l)} \cap \Omega^{(k)}|$ satisfies the following hyper-geometric distribution:

$$\mathbb{P}\left(\left|\Omega^{(l)} \cap \Omega^{(k)}\right| = t\right) = \frac{\binom{s}{t}\binom{p-s}{s-t}}{\binom{p}{s}}, \quad t = 0, \ldots, s.$$

Let $Z_{kl} = \left|\Omega^{(l)} \cap \Omega^{(k)}\right|$. Then for any $s/2 \leq t \leq s$,

$$\begin{aligned} \mathbb{P}(Z = t) = & \frac{\frac{s \cdots (s-t+1)}{t!} \cdot \frac{(p-s)\cdots(p-2s+t+1)}{(s-t)!}}{\frac{p \cdots (p-s+1)}{s!}} \leq \binom{s}{t} \cdot \left(\frac{s}{p-s+1}\right)^t \\ \leq & 2^s \left(\frac{s}{p-s+1}\right)^t \leq \left(\frac{4s}{p-s+1}\right)^t. \end{aligned} \tag{160}$$

Next, by Bernstein's inequality,

$$\begin{aligned} & \mathbb{P}\left(\left\|\mathbf{A}^{(k)} - \mathbf{A}^{(l)}\right\|_{1,1} \leq sr/2 \middle| Z\right) \overset{(159)}{=} \mathbb{P}\left(\text{Bin}\left(rZ, 1/2\right) \geq 3sr/4 \middle| Z\right) \\ = & \mathbb{P}\left(2\text{Bin}(rZ, 1/2) - rZ \geq \frac{3sr}{2} - rZ\right) \\ \leq & \begin{cases} 2\exp\left(-\frac{(3sr/2 - Zr)^2}{rZ + (3sr/2 - Zr)/3}\right), & s/2 \leq Z \leq s; \\ 0, & Z < s/2. \end{cases} \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}\left(\left\|\mathbf{A}^{(k)}-\mathbf{A}^{(l)}\right\|_{1,1}\leq sr/2\right)\leq&\sum_{s/2\leq t\leq s}\mathbb{P}\left(\left\|\mathbf{A}^{(k)}-\mathbf{A}^{(l)}\right\|_{1,1}\leq sr/2\middle|Z=t\right)\cdot\mathbb{P}\left(Z=t\right)\\ \leq&\sum_{s/2\leq t\leq s}2\exp\left(-\frac{(3sr/2-tr)^2}{rt+(3sr/2-tr)/3}\right)\left(\frac{4s}{p-s+1}\right)^t\\ \leq&\sum_{s/2\leq t\leq s}2\exp\left(-\frac{(3sr/2-sr)^2}{sr+(3sr/2-sr)/3}\right)\left(\frac{4s}{p-s+1}\right)^t\\ \leq&\sum_{t\geq s/2}2\exp\left(-sr/14\right)\cdot\left(4s/(p-r+1)\right)^t\\ \leq&2\exp(-sr/14)2\cdot(4s/(p-s+1))^{s/2}\\ \leq&4\exp\left(-c(sr+s\log(ep/s))\right)\end{aligned}$$

for some uniform constant $c>0$. Finally,

$$\begin{aligned}&\mathbb{P}\left(\forall 1\leq k\neq l\leq N,\left\|\mathbf{A}^{(k)}-\mathbf{A}^{(l)}\right\|_{1,1}>sr/2\right)\\ \geq&1-\binom{N}{2}\mathbb{P}\left(\left\|\mathbf{A}^{(k)}-\mathbf{A}^{(l)}\right\|_{1,1}\leq sr/2\right)\geq 1-\frac{N^2}{2}\cdot 4\exp\left(-c(sr+s\log(ep/s))\right)\end{aligned}$$

We can see if $N\leq\exp(c(sr+s\log(ep/s)))$ for some uniform constant $c>0$, the previous event happens with a positive probability, which means there exists fixed $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}$ satisfying the targeting condition (158) for some $N\geq\exp(c(sr+s\log(p/s)))$. $\square$