Optimal High-Order Tensor SVD via Tensor-Train Orthogonal Iteration

Yuchen Zhou, Anru R. Zhang, Lili Zheng, and Yazhen Wang

***Abstract*— This paper studies a general framework for high-order tensor SVD. We propose a new computationally efficient algorithm, tensor-train orthogonal iteration (TTOI), that aims to estimate the low tensor-train rank structure from the noisy high-order tensor observation. The proposed TTOI consists of initialization via TT-SVD [Oseledets (2011)] and new iterative backward/forward updates. We develop the general upper bound on estimation error for TTOI with the support of several new representation lemmas on tensor matricizations. By developing a matching information-theoretic lower bound, we also prove that TTOI achieves the minimax optimality under the spiked tensor model. The merits of the proposed TTOI are illustrated through applications to estimation and dimension reduction of high-order Markov processes, numerical studies, and a real data example on New York City taxi travel records. The software of the proposed algorithm is available online (https://github.com/Lili-Zheng-stat/TTOI).**

***Index Terms*— Tensor SVD, tensor-train, high-order tensors, orthogonal iteration, minimax optimality, high-order Markov chain.**

I. Introduction

TENSORS, or high-order arrays, have attracted increasing attention in modern machine learning, computational mathematics, statistics, and data science. Some specific examples include recommender systems [1], [2], neuroimaging analysis [3], [4], latent variable learning [5], multidimensional convolution [6], signal processing [7], neural network [8], [9], computational imaging [10], [11], contingency table [12], [13]. In addition to low-order tensors (e.g., tensor with a relatively small value of order number), the high-order tensors also commonly arise in applications in statistics and machine learning. For example, in convolutional neural networks, parameters in fully connected layers can be represented as high-order tensors [14], [15]. In an order-d Markov process, where the future states depend on jointly the current and $(d-1)$ previous states, the transition probabilities form an order-$(d+1)$ tensor. For an order-d Markov decision process, the transition probabilities can be represented by an order-$(2d+1)$ tensor, with additional d directions representing past d actions. High-order tensors are also used to represent the joint probability in Markov random fields [16].

Compared to the low-order tensors, high-order tensors encompass much more parameters and sophisticated structure, while leading to inhibitive cost in storage, processing, and analysis: an order-d dimension-p tensor contains p^d parameters. To address this issue, some low-dimensional parametrization is usually considered to capture the most informative subspaces in the tensor. In particular, the tensor-train (TT) decomposition [17]–[21] introduced a classic low-dimensional parameterization to model the subspaces and latent cores in high-order tensor structures. TT decomposition has been used in a wide range of applications in physics and quantum computation [18], [21]–[24], signal processing [7], and supervised learning [25] among many others. For example, the TT decomposition framework is utilized in quantum information science for modeling complex quantum states and handling the quantum mean value problem [18], [21]–[23]. The TT-decomposition of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is defined as below:

$$\begin{aligned}&\boldsymbol{\mathcal{X}}_{i_1,\cdots,i_d}\\ =&G_{1,[i_1,:]}\boldsymbol{\mathcal{G}}_{2,[:,i_2,:]}\cdots\boldsymbol{\mathcal{G}}_{d-1,[:,i_{d-1},:]}G^\top_{d,[i_d,:]}\\ =&\sum_{\alpha_1=1}^{r_1}\cdots\sum_{\alpha_{d-1}=1}^{r_{d-1}}G_{1,[i_1,\alpha_1]}\boldsymbol{\mathcal{G}}_{2,[\alpha_1,i_2,\alpha_2]}\cdots\\ &\qquad\boldsymbol{\mathcal{G}}_{d-1,[\alpha_{d-2},i_{d-1},\alpha_{d-1}]}G_{d,[i_d,\alpha_{d-1}]}.\end{aligned}\qquad(1)$$

Here, the smallest values of $r_1,\ldots,r_{d-1}$ that enable the decomposition (1) are called the *TT-rank* of $\boldsymbol{\mathcal{X}}$. Reference [17] shows that the TT-rank $r_k = \mathrm{rank}([\boldsymbol{\mathcal{X}}]_k)$, i.e., the rank of the kth sequential unfolding of $\boldsymbol{\mathcal{X}}$ (see formal definition of sequential unfolding in Section II-A). $G_1 \in \mathbb{R}^{p_1\times r_1}$, $\boldsymbol{\mathcal{G}}_k \in \mathbb{R}^{r_{k-1}\times p_k\times r_k}$, $G_d \in \mathbb{R}^{p_d\times r_{d-1}}$ are the *TT-cores* that multiply

Manuscript received October 23, 2020; revised November 7, 2021; accepted January 14, 2022. Date of publication February 18, 2022; date of current version May 20, 2022. The work of Yuchen Zhou and Anru R. Zhang was supported in part by the National Science Foundation (NSF) under Grant CAREER-1944904 and Grant DMS-1811868, and in part by the National Institutes of Health (NIH) under Grant R01GM131399. The work of Yazhen Wang was supported in part by NSF under Grant DMS-1707605 and Grant DMS-1913149. *(Corresponding author: Anru R. Zhang.)*

Yuchen Zhou was with the Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 USA. He is now with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: yczhou@wharton.upenn.edu).

Anru R. Zhang was with the Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 USA. He is now with the Department of Biostatistics and Bioinformatics, the Department of Computer Science, the Department of Mathematics, and the Department of Statistical Science, Duke University, Durham, NC 27710 USA (e-mail: anru.zhang@duke.edu).

Lili Zheng was with the Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 USA. She is now with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: lz67@rice.edu).

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

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3152733.

Digital Object Identifier 10.1109/TIT.2022.3152733

sequentially like a "train": $\mathcal{X}_{i_1,\cdots,i_d}$ equals the product of i_1th vector in G_1, i_2th matrix in $\mathcal{G}_2$, ..., i_{d-1}th matrix in $\mathcal{G}_{d-1}$, and i_dth vector in G_d. For convenience of presentation, we simplify (1) to

$$\mathcal{X} = [\![G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d]\!]$$

and denote $r_0 = r_d = 1$ throughout the paper. In particular, the TT rank and TT decomposition reduce to the regular matrix rank and decomposition when $d = 2$. If all dimensions p and ranks r are the same, the TT-parametrization involves $O(2pr + (d-2)pr^2)$ values, which can be significantly smaller than the ones for Tucker-decomposition $O(r^d + dpr)$ and the regular parameterization $O(p^d)$.

In most of the existing literature, the TT-decomposition was considered under the deterministic settings, and the central goal was often to *approximate* the nonrandom high-order tensors by low-dimensional structures [17], [26], [27]. However, in modern applications in data science such as Markov processes, Markov decision processes, and Markov random fields, the (transition) probability tensor computed based on data is often a random realization of the underlying true tensor. In these cases, the *estimation* of the underlying low-dimensional parameters hidden in the noisy observations can be more important: an accurate estimation of the transition tensor renders reliable prediction for future states in high-order Markov chains and better decision-making in high-order Markov decision processes; an accurate estimation of probability tensor sheds light on the underlying relationship among different variables in a random system [16]. To achieve such a goal, it is crucial to develop dimension reduction methods that can incorporate TT-decomposition into probabilistic models. Since singular value decomposition (SVD) is one of the most important dimension reduction methods involving probabilistic models for matrices, and there is no counterpart of it for high-order tensors, we aim to fill this void by developing a statistical framework and a computationally feasible method for high-order tensor SVD in this paper.

A. Problem Formulation

This paper focuses on the following *high-order tensor SVD model*. Suppose we observe an order-d tensor $\mathcal{Y}$ that contains a hidden tensor-train (TT) low-rank structure:

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z}, \quad \mathcal{Y}, \mathcal{X}, \mathcal{Z} \in \mathbb{R}^{\otimes_{k=1}^{d} p_k}. \tag{2}$$

Here, $\mathcal{X}$ is TT-decomposable as (1) and $\mathcal{Z}$ is a noise tensor. Our goal is to estimate $\mathcal{X}$ and the TT cores of $\mathcal{X}$ based on $\mathcal{Y}$. To this end, a straightforward idea is to minimize the approximation error as follows,

$$\widehat{\mathcal{X}} = \mathop{\arg\min}_{\mathcal{A} \text{ is decomposable as (1)}} \|\mathcal{Y} - \mathcal{A}\|_{\mathrm{F}}^2. \tag{3}$$

However, the approximation error minimization (3) is highly non-convex and finding the global optimal solution, even if the rank $r_1 = \cdots = r_{d-1} = 1$, is NP-hard in general [28]. Instead, a variety of computationally feasible methods have been proposed to approximate the best tensor-train low-rank decomposition in the literature. TT-SVD, a sequential singular value thresholding scheme, was introduced by [17] to be discussed in detail later. Reference [17] also proposed TT-rounding via sequential QR decompositions, which reduces the TT-rank while ensuring approximation accuracy. Reference [29] introduced the alternating minimal energy algorithm to reconstruct a TT-low-rank tensor approximately based on only a small proportion of revealed entries of the target tensor. [30, Section L.2] proposed a sketching-based algorithm for fast low TT rank approximation of arbitrary tensors. Reference [26] studied the tensor-train decomposition for functional tensors. Reference [31] proposed the FastTT algorithm for fast sparse tensor decomposition based on parallel vector rounding and TT-rounding. Reference [32] studied dynamical approximation with TT format for time-dependent tensors. Reference [33] proposed the alternating least squares for tensor completion in the TT format. Reference [34] studied the completion of low TT rank tensor and the applications to color image and video recovery. Reference [35] studied the Riemannian optimization methods for TT decomposition and completion. Also see [36] for a TT decomposition library in TensorFlow. To our best knowledge, the estimation performance of most procedures here remains unclear. Departing from these existing work, in this paper, we make a first attempt to minimize the estimation error of $\mathcal{X}$ in addition to achieving the minimal approximation error under possibly random settings.

B. Our Contributions

Under Model (2), we make the following contributions to high-order tensor SVD in this paper.

First, we propose a new algorithm, *Tensor-Train Orthogonal Iteration* (TTOI), that provides a computationally efficient estimation of the low-rank TT structure from the noisy observation. The proposed algorithm includes two major steps. First, we obtain initial estimates $\widehat{G}_1^{(0)}, \widehat{\mathcal{G}}_2^{(0)}, \ldots, \widehat{\mathcal{G}}_{d-1}^{(0)}, \widehat{G}_d$ by performing forward sequential SVD based on matricizations and projections. This step was known as TT-SVD in the literature [17]. Next, we utilize the initialization and perform the newly developed *backward updates* and *forward updates* alternatively and iteratively. The TTOI procedure will be discussed in detail in Section II.

To see why the TTOI iterations yield better estimation than the classic TT-SVD method, recall that TT-SVD first performs singular value thresholding on $[\mathcal{Y}]_1$, i.e., the unfolding of $\mathcal{Y}$, without any additional updates (see detailed procedure of TT-SVD and formal definition of $[\mathcal{Y}]_1$ in Section II-A), which can be inaccurate since $[\mathcal{Y}]_1$, a p_1-by-$\prod_{k=2}^{d} p_k$ matrix, has a great number of columns. In contrast, TTOI iteration utilizes the intermediate outcome of the previous iteration to substantially reduce the dimension of $[\mathcal{Y}]_1$ while performing singular value thresholding. In Figure 1, we provide a simple simulation example to show that even one TTOI iteration can significantly improve the estimation of the left singular subspace of G_1 (left panel) and the overall tensor $\mathcal{X}$ (right panel). Therefore, a one-step TTOI, i.e., the initialization with one TTOI iteration, can be used in practice when the computational cost is a concern.

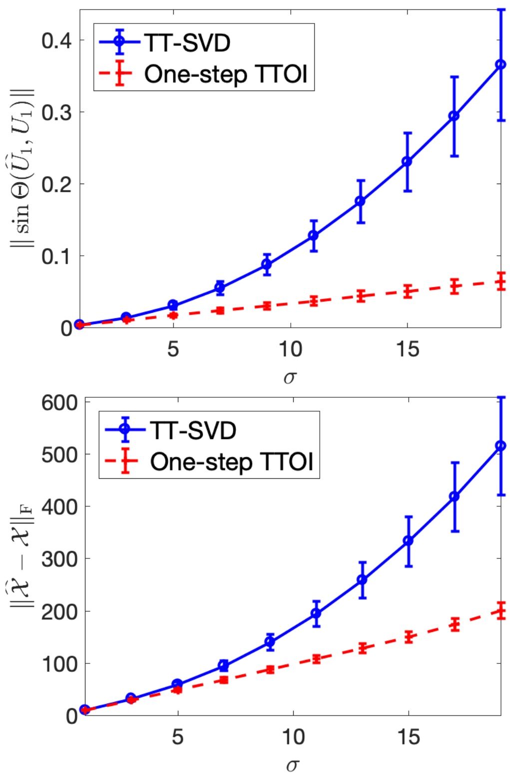


Fig. 1. Average estimation error (dots) and standard deviation (bars) of $\|\sin\Theta(\widehat{U}_1, U_1)\|$ and $\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_{\rm F}$ by TT-SVD and one-step TTOI. Both algorithms are performed based on the observation $\boldsymbol{\mathcal{Y}}$ generated from (2), where $\boldsymbol{\mathcal{Z}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $\boldsymbol{\mathcal{X}}$ is a randomly generated order-5 tensor based on (1) with $p = 20, r = 1$, $G_1, \boldsymbol{\mathcal{G}}_2, \ldots, \boldsymbol{\mathcal{G}}_{d-1}, G_d \overset{\text{i.i.d.}}{\sim} N(0, 1)$.

We develop theoretical guarantees for TTOI. In particular, we introduce a series of representation lemmas for tensor matricizations with TT format. Based on them, we develop a deterministic upper bound of estimation error for both forward and backward updates in TTOI iterations. Under the benchmark setting of spiked tensor model, we develop matching upper/lower bounds and prove that the proposed TTOI algorithm achieves the minimax optimal rate of estimation error. To the best of our knowledge, this is the first statistical optimality results for high-order tensors with TT format. We also prove for any high-order tensor, TTOI iteration has monotone decreasing approximation error with respect to the iteration index.

Moreover, to break the curse of dimensionality in high-order Markov processes, we study the state aggregatable high-order Markov processes and establish a key connection to TT decomposable tensors. We propose a TTOI estimator for the transition probability tensor in high-order state-aggregatable Markov processes and establish the theoretical guarantee. We conduct simulation experiments to demonstrate the performance of TTOI and validate our theoretical findings. We also apply our method to analyze a New York taxi dataset. By modeling taxi trips as trajectories realized from a citywide Markov chain, we found that the Manhattan traffic zone exhibits high-order Markovian dependence and the proposed TTOI reveals latent traffic patterns and meaningful partition of Manhattan traffic zones. Finally, we discuss several applications that our proposed algorithm is applicable to, including transition probability tensor estimation in high-order Markov decision processes and joint probability tensor estimation in Markov random fields.

C. Related Literature

In addition to the aforementioned literature on TT decomposition, our work is also related to a substantial body of work on matrix/tensor decomposition and SVD, spiked tensor model, etc. These literature are from a range of communities including applied mathematics, information theory, machine learning, scientific computing, signal processing, and statistics. Here we try to review existing literature in these communities without claiming this literature survey is exhaustive.

First, the matrix singular value thresholding was commonly used and extensively studied in various problems in data science, including matrix denoising [37]–[39], matrix completion [40]–[43], principal component analysis (PCA) [44], Markov chain state aggregation [45]. Such the task was also widely considered for tensors of order-3 or higher. In particular, to perform SVD and decomposition for tensors with Tucker low-rank structures, [46], [47] introduced the higher-order SVD (HOSVD) and higher-order orthogonal iteration (HOOI). Reference [48] established the statistical and computational limits of tensor SVD, compared the theoretical properties of HOSVD and HOOI, and proved that HOOI achieves both statistical and computational optimality. Reference [49] introduced the sequentially truncated higher-order singular value decomposition (ST-HOSVD). Reference [50] introduced a thresholding & projection based algorithm for sparse tensor SVD. A non-exhaustive list of methods for SVD and decomposition for tensors with CP low-rank structures include alternating least squares [51], [52], eigendecomposition-based approach [53], enhanced line search [54], power iteration with SVD-based initialization [5], simultaneous diagonalization and higher-order SVD [55].

In addition, the spiked tensor model and tensor principal component analysis (tensor PCA) are widely discussed in the literature. References [56]–[61] considered the statistical and computational limits of rank-1 spiked tensor model. Reference [62] studied the statistical and computational phase transitions and theoretical properties of the approximate message passing algorithm (AMP) under a Bayesian spiked tensor model. References [63] and [64] developed the regularization-based methods for tensor PCA. References [65]–[68] studied the robust tensor PCA to handle the possible outliers from the tensor observation.

Different from Tucker and CP decompositions, which have been a pinpoint in the enormous existing literature on tensors, we focus on the TT-structure associated with high-order tensors for the following reasons: (1) Tucker and CP decompositions do not involve the sequential structure of different modes, i.e., the Tucker and CP decompositions still hold if the d modes are arbitrarily permuted. While in applications such as high-order Markov process, high-order Markov decision process, and fully connected layers of deep neural networks, the order of different modes can be crucial; (2) the number of entries involved in the low-Tucker-rank parameterization grows exponentially with respect to the order d (r^d); (3) methods that explore CP low-rank structure can be numerically unstable for high-order tensors in computation as pointed out by [27]. In comparison, the TT-structure incorporates the order of different modes sequentially and involves much fewer parameters for high-order tensors, which renders it more suitable in many scenarios.

In Section V, we will further discuss the application of TTOI on high-order Markov processes and state aggregation. This problem is related to a body of literature on dimension reduction and state aggregation for Markov processes that we will discuss in Section V.

D. Organization

The rest of the article is organized as follows. In Section II, after a brief introduction of the notation and preliminaries, we introduce the procedure of the tensor-train orthogonal iteration. The theoretical results, including three representation lemmas, a general estimation error bound, and the minimax optimal upper and lower bounds under the spiked tensor model, are provided in Sections III and IV. The application to high-order Markov chains is discussed in Section V. The simulation and real data analysis are provided in Sections VI-A and VI-B, respectively. Discussions and further applications to Markov random fields and high-order Markov decision processes are briefly discussed in Section VII. All technical proofs are provided in Section A.

II. Procedure of Tensor-Train Orthogonal Iteration

A. Notation and Preliminaries

We first introduce the notation and preliminaries to be used throughout the paper. We use the lowercase letters, e.g., x, y, z, to denote scalars or vectors. We use $C, c, C_0, c_0, \dots$ to denote generic constants, whose actual values may change from line to line. A random variable z is σ-sub-Gaussian if $\mathbb{E}e^{t(z-\mathbb{E}z)} \leq e^{\sigma^2 t^2/2}$ for any $t \in \mathbb{R}$. We say $a \lesssim b$ or $a = O(b)$ if $a \leq Cb$ for some uniform constant $C > 0$. We write $a = \widetilde{O}(b)$ if $a = O(b \log^{C'}(b))$ for constant $C' > 0$. The capital letters, e.g., X, Y, Z, are used to denote matrices. Specifically, $\mathbb{O}_{p,r} := \{U \in \mathbb{R}^{p\times r} : U^\top U = I_r\}$ is the set of all p-by-r matrices with orthogonal columns. For $U \in \mathbb{O}_{p,r}$, let $U_\perp \in \mathbb{O}_{p,p-r}$ be the orthonormal complement of U, and let $P_U = UU^\top$ denote the projection matrix onto the column space of U. For any matrix $A \in \mathbb{R}^{p_1\times p_2}$, let $A = \sum_{i=1}^{p_1\wedge p_2} s_i u_i v_i^\top$ be the singular value decomposition, where $s_1(A) \geq \cdots \geq s_{p_1\wedge p_2}(A) \geq 0$ are the singular values of A in non-increasing order. Define $s_{\min}(A) = s_{p_1\wedge p_2}(A)$, $\text{SVD}_r^L(A) = [u_1 \dots u_r] \in \mathbb{O}_{p_1,r}$, and $\text{SVD}_r^R(A) = [v_1 \dots v_r] \in \mathbb{O}_{p_2,r}$ be the smallest non-trivial singular value, leading r left singular vectors, and leading r right singular vectors of A, respectively. We also write $\text{SVD}^L(A) = \text{SVD}^L_{p_1\wedge p_2}(A)$ and $\text{SVD}^R(A) = \text{SVD}^L_{p_1\wedge p_2}(A)$ as the collection of all left and right singular vectors of A, respectively. Define the Frobenius and spectral norms of A as $\|A\|_{\text{F}} = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} A_{ij}^2} = \sqrt{\sum_{i=1}^{p_1\wedge p_2} s_i^2(A)}$ and $\|A\| = s_1(A) = \max_{x\in\mathbb{R}^{p_2}} \|Ax\|_2/\|x\|_2$. For any two matrices $U \in \mathbb{R}^{m_1\times n_1}$ and $V \in \mathbb{R}^{m_2\times n_2}$, let

$$U \otimes V = \begin{bmatrix} U_{11}\cdot V & \dots & U_{1n_1}\cdot V \\ \vdots & & \vdots \\ U_{m_1 1}\cdot V & \dots & U_{m_1 n_1}\cdot V \end{bmatrix} \in \mathbb{R}^{(m_1 m_2)\times(n_1 n_2)}$$

be their Kronecker product. To quantify the distance among subspaces, we define the principle angles between $U, \widehat{U} \in \mathbb{O}_{p,r}$ as an r-by-r diagonal matrix: $\Theta(U,\widehat{U}) = \text{diag}(\arccos(s_1), \dots, \arccos(s_r))$, where $s_1 \geq \cdots \geq s_r \geq 0$ are the singular values of $U^\top\widehat{U}$. Define the sinΘ norm as

$$\begin{aligned} &\|\sin\Theta(U,\widehat{U})\| \\ =&\|\text{diag}\left(\sin(\arccos(s_1)), \dots, \sin(\arccos(s_r))\right)\| \\ =&\sqrt{1-s_r^2}. \end{aligned}$$

The boldface calligraphic letters, e.g., $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}}$, are used to denote tensors. For an order-d tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{\otimes_{i=1}^d p_i}$ and $1 \leq k \leq d-1$, we define $[\boldsymbol{\mathcal{X}}]_k \in \mathbb{R}^{(p_1\times\cdots\times p_k)\times(p_{k+1}\cdots p_d)}$ as the *sequential unfolding* of $\boldsymbol{\mathcal{X}}$ with rows enumerating all indices in Modes $1, \dots, k$ and columns enumerating all indices in Modes $(k+1), \cdots, d$, respectively. That is, for any $1 \leq k \leq d$ and $1 \leq i_k \leq p_k$,

$$([\boldsymbol{\mathcal{X}}]_k)_{\xi_1(i_1,\dots,i_d;k),\xi_2(i_1,\dots,i_d;k)} = \boldsymbol{\mathcal{X}}_{i_1\dots i_d},$$

where $\xi_1(i_1, \dots, i_d; k) = (i_k - 1)p_1\cdots p_{k-1} + (i_{k-1} - 1)p_1\cdots p_{k-2} + \cdots + i_1$ and $\xi_2(i_1, \dots, i_d; k) = (i_d - 1)p_{k+1}\cdots p_{d-1} + (i_{d-1} - 1)p_{k+1}\cdots p_{d-2} + \cdots + i_{k+1}$. Following the convention of `reshape` function in MATLAB, we define the reshape of any matrix X of dimension $p_1\cdots p_k \times p_{k+1}\cdots p_d$ as an inverse operation of tensor matricization: $\boldsymbol{\mathcal{X}} = \text{Reshape}(X, p_1, p_2, \dots, p_d)$ if $X = [\boldsymbol{\mathcal{X}}]_k$. For any two matrices $A \in \mathbb{R}^{q_1\times q_2 q_3}$ and $\widetilde{A} \in \mathbb{R}^{q_1 q_2\times q_3}$, we denote $\widetilde{A} = \text{Reshape}(A, q_1q_2, q_3)$ and $A = \text{Reshape}(\widetilde{A}, q_1, q_2q_3)$ if and only if

$$\widetilde{A}_{(i_2-1)p_1+i_1,i_3} = A_{i_1,(i_3-1)p_2+i_2}, \quad \forall 1 \leq i_j \leq q_j, j = 1, 2, 3.$$

We also define the tensor Frobenius norm of $\boldsymbol{\mathcal{X}}$ as $\|\boldsymbol{\mathcal{X}}\|_{\text{F}}^2 = \sum_{i_1=1}^{p_1}\cdots\sum_{i_d=1}^{p_d} \boldsymbol{\mathcal{X}}_{i_1,\dots,i_d}^2$. For any matrix $A \in \mathbb{R}^{p_1\times p_2}$ and any tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$, let $\text{vec}(A)$ and $\text{vec}(\boldsymbol{\mathcal{B}})$ be the vectorization of A and $\boldsymbol{\mathcal{B}}$, respectively. Formally, for any $1 \leq k \leq d$ and $1 \leq i_k \leq p_k$,

$$(\text{vec}(\boldsymbol{\mathcal{B}}))_{(i_d-1)p_1\cdots p_{d-1}+(i_{d-1}-1)p_1\cdots p_{d-2}+\cdots+i_1} = \boldsymbol{\mathcal{B}}_{i_1,\dots,i_d}.$$

B. Procedure of Tensor-Train Orthogonal Iteration

We are now in position to introduce the procedure of Tensor-Train Orthogonal Iteration (TTOI). The pseudocode of the overall procedure is given in Algorithm 1. TTOI includes three main parts: we first run *initialization*, then perform *backward update* and *forward update* alternatively and iteratively.

- **Part 1: Initialization.** First, we obtain an initial estimate of TT-cores $G_1, \boldsymbol{\mathcal{G}}_2, \dots, \boldsymbol{\mathcal{G}}_{d-1}, G_d$. This step is the tensor-train-singular value decomposition (TT-SVD) originally introduced by [17].

Algorithm 1 Tensor-Train Orthogonal Iteration (TTOI)

Input: $\boldsymbol{\mathcal{Y}}, \{p_k\}_{k=1}^d, \{r_k\}_{k=1}^{d-1}$, increment tolerance $\varepsilon > 0$, maximum number of iterations $t_{\max}$

Obtain Initialization $\widetilde{R}_1^{(0)}, \ldots, \widetilde{R}_{d-1}^{(0)}, \widehat{\boldsymbol{\mathcal{X}}}^{(0)}$ by Algorithm 1(a)
for $t = 1, \ldots, t_{\max}$ **do**
  if t is odd **then**
    Apply Algorithm 1(b) with input $\widetilde{R}_1^{(t-1)}, \ldots, \widetilde{R}_{d-1}^{(t-1)}$ to obtain $\widehat{V}_1^{(t)}, \ldots, \widehat{V}_d^{(t)}, \widehat{\boldsymbol{\mathcal{X}}}^{(t)}$
  else
    Apply Algorithm 1(c) with input $\widehat{V}_1^{(t-1)}, \ldots, \widehat{V}_d^{(t-1)}$ to obtain $\widetilde{R}_1^{(t)}, \ldots, \widetilde{R}_{d-1}^{(t)}, \widehat{\boldsymbol{\mathcal{X}}}^{(t)}$
  end if
  If $\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{\mathcal{X}}}^{(t-1)}\|_{\mathrm{F}}^2 \leq \varepsilon$ **then** break from the for loop
end for

Output: $\widehat{\boldsymbol{\mathcal{X}}} = \widehat{\boldsymbol{\mathcal{X}}}^{(t)}$

Algorithm 1(a) Initialization (TT-SVD [17])

Input: $\boldsymbol{\mathcal{Y}}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^d$

Calculate $R_1^{(0)} = [\boldsymbol{\mathcal{Y}}]_1$
for $k = 1, \ldots, d-1$ **do**
  $\widehat{U}_k^{(0)} = \mathrm{SVD}_{r_k}^L(R_k^{(0)})$
  If $k = 1$ **then** $U_{\text{prod}}^{(0)} = \widehat{U}_k^{(0)}$ **else** $U_{\text{prod}}^{(0)} = (I_{p_k} \otimes U_{\text{prod}}^{(0)})\widehat{U}_k^{(0)}$
  $\widetilde{R}_k^{(0)} = \widehat{U}_k^{(0)\top} R_k^{(0)}$
  If $k < d-1$ **then** $R_{k+1}^{(0)} = \mathrm{reshape}(\widetilde{R}_k^{(0)}, r_k p_{k+1}, p_{k+2} \cdots p_d)$
end for
$[\widehat{X}^{(0)}]_{d-1} = U_{\text{prod}}^{(0)} \widetilde{R}_{d-1}^{(0)}$
Reshape $[\widehat{X}^{(0)}]_{d-1} \in \mathbb{R}^{(p_1 \cdots p_{d-1}) \times p_d}$ to $\widehat{\boldsymbol{\mathcal{X}}}^{(0)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$

Output: $\widetilde{R}_1^{(0)}, \ldots, \widetilde{R}_{d-1}^{(0)}, \widehat{\boldsymbol{\mathcal{X}}}^{(0)}$

(i) Let $R_1^{(0)}$ be the unfolding of $\boldsymbol{\mathcal{Y}}$ along Mode 1. We compute the top-r_1 SVD of $R_1^{(0)}$. Let $\widehat{U}_1^{(0)} \in \mathbb{O}_{p_1, r_1}$ be the first r_1 left singular vectors of $R_1^{(0)}$ and calculate $\widetilde{R}_1^{(0)} = (\widehat{U}_1^{(0)})^\top R_1^{(0)} \in \mathbb{R}^{r_1 \times (p_2 \ldots p_d)}$. Then, $\widehat{U}_1^{(0)}$ is an initial estimate of the subspace that G_1 lies in and $\widetilde{R}_1^{(0)}$ can be seen as the projection residual.

(ii) Next, we realign the entries of $\widetilde{R}_1^{(0)} \in \mathbb{R}^{r_1 \times (p_2 \ldots p_d)}$ to $R_2^{(0)} \in \mathbb{R}^{(r_1 p_2) \times (p_3 \ldots p_d)}$, where the rows and columns of $R_2^{(0)}$ correspond to indices of Modes-1, 2 and Modes-$3, \ldots, d$, respectively. Then, we evaluate the top-r_2 SVD of $R_2^{(0)}$. Let $\widehat{U}_2^{(0)}$ be the first r_2 left singular vectors of $R_2^{(0)}$ and evaluate $\widetilde{R}_2^{(0)} = (\widehat{U}_2^{(0)})^\top R_2^{(0)} \in \mathbb{R}^{r_2 \times p_3 \cdots p_d}$. Again, $\widehat{U}_2^{(0)}$ is an estimate of the singular subspace that $\mathcal{G}_2$ lies on and $\widetilde{R}_2^{(0)}$ is the projection residual for the next calculation.

(iii) We apply Step (ii) on $\widetilde{R}_2^{(0)}$ to obtain $\widehat{U}_3^{(0)} \in \mathbb{O}_{r_2 p_3, r_3}$ and $\widetilde{R}_3^{(0)} \in \mathbb{R}^{r_3 \times (p_4 \cdots p_d)}$; …; apply Step (ii) on $\widetilde{R}_{d-2}^{(0)}$ to obtain $\widehat{U}_{d-1}^{(0)} \in \mathbb{O}_{r_{d-2} p_{d-1}, r_{d-1}}$ and $\widetilde{R}_{d-1}^{(0)} \in \mathbb{R}^{r_{d-1} \times p_d}$. Then we reshape matrix $\widehat{U}_k^{(0)} \in \mathbb{R}^{(p_k r_{k-1}) \times r_k}$ to tensor $\widehat{\boldsymbol{\mathcal{U}}}_k^{(0)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$ for $k = 2, \ldots, d-1$. Now, $\left(\widehat{U}_1^{(0)}, \widehat{\boldsymbol{\mathcal{U}}}_2^{(0)}, \ldots, \widehat{\boldsymbol{\mathcal{U}}}_{d-1}^{(0)}, \widetilde{R}_{d-1}^{(0)\top}\right)$ yield the initial estimates of TT-cores of $\boldsymbol{\mathcal{X}}$ and we expect that

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{X}}^{(0)} = [\![\widehat{U}_1^{(0)}, \widehat{\boldsymbol{\mathcal{U}}}_2^{(0)}, \cdots, \widehat{\boldsymbol{\mathcal{U}}}_{d-1}^{(0)}, \widetilde{R}_{d-1}^{(0)}]\!].$$

The initialization step is summarized to Algorithm 1(a) and illustrated in Figure 2. In summary, we perform SVD on some "residual" $R_k^{(0)}$ sequentially for $k = 1, \ldots, d-1$. As will be shown in Lemma 3, $R_k^{(0)}$ satisfies

$$R_k^{(0)} = (I_{p_k} \otimes \widehat{U}_{k-1}^{(0)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(0)\top})[\boldsymbol{\mathcal{Y}}]_k,$$

where $[\boldsymbol{\mathcal{Y}}]_k \in \mathbb{R}^{(p_1 \cdots p_k) \times (p_{k+1} \cdots p_d)}$ is the kth sequential unfolding of $\boldsymbol{\mathcal{Y}}$ (see definition in Section II-A). This quantity plays a key role in the backward update next.

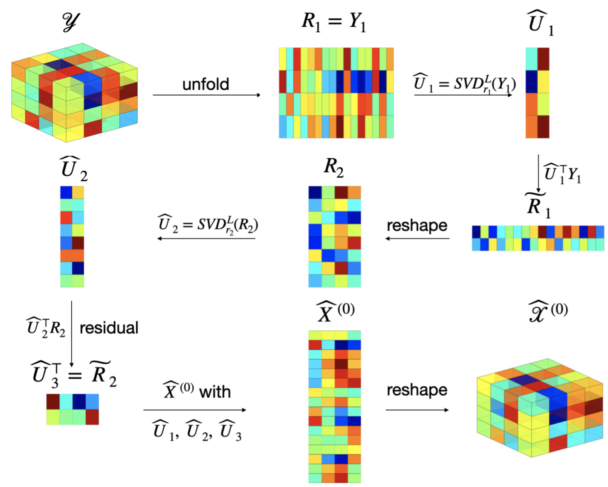


Fig. 2. A pictorial illustration of initialization (Algorithm 1(a), $d = 3$).

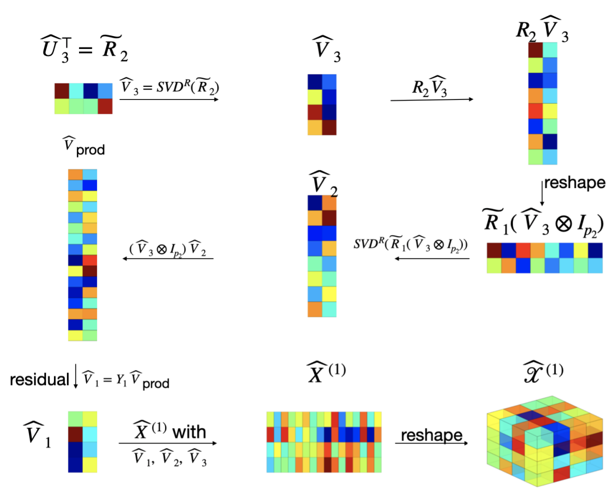


Fig. 3. A pictorial illustration of TT-backward update (Algorithm 1(b), $d = 3$).

The initialization step mainly focuses on the left singular spaces of $[\boldsymbol{\mathcal{X}}]_k$ while ignoring the information included in the right singular spaces. Due to this fact, we develop the

Algorithm 1(b) TT-Backward Update

Input: $\mathcal{Y}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^{d}, \widetilde{R}_1^{(t-1)}, \ldots, \widetilde{R}_{d-1}^{(t-1)}$ for odd iteration number t

for $k = 1, \ldots, d-1$ **do**
  if $k = 1$ **then**
    $\widehat{V}_{d-k+1}^{(t)} = \mathrm{SVD}^{R}_{r_{d-k}}\left(\widetilde{R}_{d-k}^{(t-1)}\right), \quad V_{\text{prod}}^{(t)} = \widehat{V}_{d-k+1}^{(t)}$
  else
    $\widehat{V}_{d-k+1}^{(t)} = \mathrm{SVD}^{R}_{r_{d-k}}\left(\widetilde{R}_{d-k}^{(t-1)}(V_{\text{prod}}^{(t)} \otimes I_{p_{d-k+1}})\right)$, $V_{\text{prod}}^{(t)} = (V_{\text{prod}}^{(t)} \otimes I_{p_{d-k+1}})\widehat{V}_{d-k+1}^{(t)}$
  end if
end for
$\widehat{V}_1^{(t)} = [\mathcal{Y}]_1 V_{\text{prod}}^{(t)}$, $[\widehat{X}^{(t)}]_1 = \widehat{V}_1^{(t)} V_{\text{prod}}^{(t)\top}$, reshape $[\widehat{X}^{(t)}]_1 \in \mathbb{R}^{p_1 \times (p_2 \cdots p_d)}$ to $\widehat{\mathcal{X}}^{(t)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$

Output: $\widehat{V}_1^{(t)}, \ldots, \widehat{V}_d^{(t)}, \widehat{\mathcal{X}}^{(t)}$

Algorithm 1(c) TT-Forward Update

Input: $\mathcal{Y}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^{d}, \widehat{V}_1^{(t-1)}, \ldots, \widehat{V}_d^{(t-1)}$ for even iteration number t

$R_1^{(t)} = [\mathcal{Y}]_1$
for $k = 1, \ldots, d-1$ **do**
  if $k = 1$ **then**
    $\widehat{U}_1^{(t)} = \mathrm{SVD}^{L}_{r_1}\left(\widehat{V}_1^{(t)}\right), \quad U_{\text{prod}}^{(t)} = \widehat{U}_1^{(t)}$
  else

$$\widehat{U}_k^{(t)} = \mathrm{SVD}^{R}_{r_k}\left(R_k^{(t)}(\widehat{V}_d^{(t-1)} \otimes I_{p_{k+1}\ldots p_{d-1}}) \cdots (\widehat{V}_{k+2}^{(t-1)} \otimes I_{p_{k+1}})\widehat{V}_{k+1}^{(t-1)}\right)$$

    $U_{\text{prod}}^{(t)} = (I_{p_k} \otimes U_{\text{prod}}^{(t)})\widehat{U}_k^{(t)}$
  end if
  $\widetilde{R}_k^{(t)} = \widehat{U}_k^{(t)\top} R_k^{(t)}$
  If $k < d-1$ **then** $R_{k+1}^{(t)} = \text{reshape}\left(\widetilde{R}_k^{(t)}, r_k p_{k+1}, p_{k+2} \cdots p_d\right)$
end for
$[\widehat{X}^{(t)}]_{d-1} = U_{\text{prod}}^{(t)} \widetilde{R}_{d-1}^{(0)}$, reshape $[\widehat{X}^{(t)}]_{d-1} \in \mathbb{R}^{(p_1 \cdots p_{d-1}) \times p_d}$ to $\widehat{\mathcal{X}}^{(t)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$

Output: $\widetilde{R}_1^{(t)}, \ldots, \widetilde{R}_{d-1}^{(t)}, \widehat{\mathcal{X}}^{(t)}$

following new backward update that utilizes both the left and right singular space estimates from the previous step to refine our estimates. Similarly, we can also perform a forward update to further improve the outcome of backward update, and then iteratively alternate between backward and forward updates. The detailed descriptions of these two updates are presented as follows, and a further explanation is given in Remark 1.

- **Part 2: Backward update.** For iterations $t = 1, 3, 5, \ldots$, we perform backward update, i.e., to sequentially obtain $\widehat{V}_d^{(t)}, \ldots, \widehat{V}_2^{(t)}$ based on the intermediate results from the $(t-1)$st iteration (0th iteration is the initialization). The pseudocode of backward update is provided in Algorithm 1(b). The calculation in Algorithm 1(b) is equivalent to

$$\begin{aligned}\widehat{V}_d^{(t)} =& \mathrm{SVD}^R\left(\widetilde{R}_{d-1}^{(t-1)}\right),\\ \widehat{V}_k^{(t)} &\\ =&\mathrm{SVD}^R\left(\widetilde{R}_{k-1}^{(t-1)}(\widehat{V}_d^{(t)} \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1}^{(t)} \otimes I_{p_k})\right)\end{aligned}$$

for $k = d-1, \ldots, 2$, and

$$\begin{aligned}\widehat{V}_1^{(t)} =& [\mathcal{Y}]_1(\widehat{V}_d^{(t)} \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3^{(t)} \otimes I_{p_2})\widehat{V}_2^{(t)}\\ &\in \mathbb{R}^{p_1 \times r_1}.\end{aligned}$$

Here,

$$\begin{aligned}&\widetilde{R}_k^{(t-1)}\\ =&(\widehat{U}_k^{(t-1)})^\top (I_{p_k} \otimes \widehat{U}_{k-1}^{(t-1)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t-1)\top})[\mathcal{Y}]_k\end{aligned}$$

are the projection residual term in the intermediate outcome of the $(t-1)$st iteration. Then, we reshape $\widehat{V}_k^{(t)\top} \in \mathbb{R}^{r_{k-1} \times (p_k r_k)}$ to $\widehat{\mathcal{V}}_k^{(t)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$. The backward updated estimate is

$$\widehat{\mathcal{X}}^{(t)} = [\![\widehat{V}_1^{(t)}, \widehat{\mathcal{V}}_2^{(t)}, \ldots, \widehat{\mathcal{V}}_{d-1}^{(t)}, \widehat{V}_d^{(t)}]\!].$$

Remark 1 (Interpretation of Backward Update): The backward updates utilize and extract the right singular vectors of the intermediate products of the $(t-1)$st iteration,

$$\begin{aligned}&\widetilde{R}_k^{(t-1)}\\ =&(\widehat{U}_k^{(t-1)})^\top (I_{p_k} \otimes \widehat{U}_{k-1}^{(t-1)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t-1)\top})[\mathcal{Y}]_k,\end{aligned}$$

as opposed to the entire data $[\mathcal{Y}]_k$. Such a dimension reduction scheme is the key to the backward update: it can simultaneously reduce the dimension of the matrix of interest, $[\mathcal{Y}]_k$, and the noise therein, while preserving the signal strength. Different from the initialization in Step 1, the backward update utilizes the information from both the forward and backward singular subspaces of the tensor-train structure of $\mathcal{X}$. See Section III for more illustration.

- **Part 3: Forward Update.** For iteration $t = 2, 4, 6, \ldots$, we perform forward update, i.e., to sequentially obtain $\widehat{U}_1^{(t)}, \ldots, \widehat{U}_d^{(t)}$ based on the intermediate results from the $(t-1)$st iteration. Essentially, the forward update can be seen as a reversion of the backward update by flipping all modes of tensor $\mathcal{Y}$. The pseudocode of this procedure is collected in Algorithm 1(c). Recall $[\mathcal{Y}]_1(\widehat{V}_d^{(t-1)} \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3^{(t-1)} \otimes I_{p_2})\widehat{V}_2^{(t-1)}$ is the intermediate product from the $(t-1)$st update. We sequentially compute

$$\widehat{U}_1^{(t)} = \mathrm{SVD}^L\left([\mathcal{Y}]_1(\widehat{V}_d^{(t-1)} \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3^{(t-1)} \otimes I_{p_2})\widehat{V}_2^{(t-1)}\right);$$

$$\widehat{U}_k^{(t)} = \mathrm{SVD}^L\left((I_{p_k} \otimes \widehat{U}_{k-1}^{(t)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t)\top})[\mathcal{Y}]_k \cdot (\widehat{V}_d^{(t-1)} \otimes I_{p_{k+1} \ldots p_{d-1}}) \cdots (\widehat{V}_{k+2}^{(t-1)} \otimes I_{p_{k+1}})\widehat{V}_{k+1}^{(t-1)}\right)$$

for $k = 2, \ldots, d-1$, and

$$\widehat{U}_d^{(t)} = \left[(\widehat{U}_{d-1}^{(t)})^\top (I_{p_{d-1}} \otimes (\widehat{U}_{d-2}^{(t)})^\top) \cdots (I_{p_{d-1}\cdots p_2} \otimes (\widehat{U}_1^{(t)})^\top)[\boldsymbol{\mathcal{Y}}]_{d-1}\right]^\top \in \mathbb{R}^{p_d \times r_{d-1}}.$$

Reshape $\widehat{U}_k^{(t)} \in \mathbb{R}^{(p_k r_{k-1}) \times r_k}$ to $\widehat{\boldsymbol{\mathcal{U}}}_k^{(t)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$ for $k = 2, \ldots, d-1$. Then, compute

$$\widehat{\boldsymbol{\mathcal{X}}}^{(t)} = [\![\widehat{U}_1^{(t)}, \widehat{\boldsymbol{\mathcal{U}}}_2^{(t)}, \ldots, \widehat{\boldsymbol{\mathcal{U}}}_{d-1}^{(t)}, \widehat{U}_d^{(t)}]\!].$$

We will explain the algebraic schemes in the TTOI procedure through several representation lemmas in Section III-A. We will also show in Theorem 2 that the objective function $\|\boldsymbol{\mathcal{Y}} - \widehat{\boldsymbol{\mathcal{X}}}^{(t)}\|_{\mathrm{F}}^2$ is monotone decreasing with respect to the iteration index t. In the large-scale scenarios that performing iterations is beyond the capacity of computing, we can reduce the number of iterations, and even to $t_{\max} = 1$, i.e., the one-step iteration, which have often yielded sufficiently accurate estimation as we will illustrate in both theory and simulation studies. Such the phenomenon has been recently discovered for HOOI in the Tucker low-rank tensor decomposition [69].

Remark 2 (Computational and storage costs of TTOI): We consider the computational and storage costs of TTOI on the p-dimensional, rank-r, order-d, and dense tensor. Since computing the first r singular vectors of an $m \times n$ matrix via block power method requires $\widetilde{O}(mnr)$ operations, initialization costs $\widetilde{O}(p^d r)$ operations, each iteration of TTOI, including forward and backward updates, costs $O(p^d r)$. Therefore, the total number of operations of TTOI with T iterations is $\widetilde{O}(p^d r) + O(Tp^d r)$, which is not significantly more than the number of elements of the target tensor. Moreover, TTOI requires $O(p^d)$ storage cost, which is not significantly more than the storage cost of the original tensor.

III. Theoretical Analysis

This section is devoted to the theoretical analysis of the proposed procedure. For convenience, we introduce the following two abbreviations for matrix sequential products: for $M_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, 1 \le i \le d-1$ and $B_j \in \mathbb{R}^{(r_j p_j) \times r_{j-1}}, 2 \le j \le d$, we denote

$$\begin{aligned} M_{\text{prod},k}^{(L)} &= (I_{p_2\cdots p_k} \otimes M_1) \cdots (I_{p_k} \otimes M_{k-1}) M_k \\ &\in \mathbb{R}^{(p_1\cdots p_k) \times r_k}, \quad \forall 1 \le k \le d-1, \\ B_{\text{prod},k}^{(R)} &= (B_d \otimes I_{p_k\cdots p_{d-1}}) \cdots (B_{k+1} \otimes I_{p_k}) B_k \\ &\in \mathbb{R}^{(p_k\cdots p_d) \times r_{k-1}}, \quad \forall 2 \le k \le d. \end{aligned}$$

Equivalently, $M_{\text{prod},k}^{(L)}$ and $B_{\text{prod},k}^{(R)}$ can be defined sequentially as

$$\begin{aligned} M_{\text{prod},1}^{(L)} &= M_1, \\ M_{\text{prod},k+1}^{(L)} &= (I_{p_{k+1}} \otimes M_{\text{prod},k}^{(L)}) M_{k+1}, \quad 1 \le k \le d-2, \\ B_{\text{prod},d}^{(R)} &= B_d, \\ B_{\text{prod},k}^{(R)} &= (B_{\text{prod},k+1}^{(R)} \otimes I_{p_k}) B_k, \quad 2 \le k \le d-1. \end{aligned}$$

A. Representation Lemmas for High-Order Tensors

Since the computation of high-order tensors with tensor-train structures involves extensive tensor algebra, we introduce the following three lemmas on the matrix representation of high-order tensors. These lemmas play a fundamental role in the later theoretical analysis.

Lemma 1 (Representation for Sequential Matricization of TT-Decomposable Tensor): Suppose $\boldsymbol{\mathcal{X}} = [\![G_1, \boldsymbol{\mathcal{G}}_2, \ldots, \boldsymbol{\mathcal{G}}_{d-1}, G_d]\!]$. Then the sequential matricization of $\boldsymbol{\mathcal{X}}$ can be written as

$$\begin{aligned} [\boldsymbol{\mathcal{X}}]_k =& (I_{p_2\cdots p_k} \otimes G_1)(I_{p_3\cdots p_k} \otimes [\boldsymbol{\mathcal{G}}_2]_2) \cdots (I_{p_k} \otimes [\boldsymbol{\mathcal{G}}_{k-1}]_2) \\ & \cdot [\boldsymbol{\mathcal{G}}_k]_2 [\boldsymbol{\mathcal{G}}_{k+1}]_1 \left([\boldsymbol{\mathcal{G}}_{k+2}]_1 \otimes I_{p_{k+1}}\right) \cdots \\ & \left([\boldsymbol{\mathcal{G}}_{d-1}]_1 \otimes I_{p_{k+1}\cdots p_{d-2}}\right) \left(G_d^\top \otimes I_{p_{k+1}\cdots p_{d-1}}\right). \end{aligned} \quad (4)$$

Lemma 2 (Representation of Tensor Reshaping): For any tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{\otimes_{k=1}^d p_k}$ and $1 \le i < j \le d-1$, we have

$$\begin{aligned} [\boldsymbol{\mathcal{T}}]_j &= (I_{p_{i+1}\cdots p_j} \otimes [\boldsymbol{\mathcal{T}}]_i) A^{(p_{i+1}\cdots p_j, p_{j+1}\cdots p_d)}, \\ [\boldsymbol{\mathcal{T}}]_i &= A^{(p_{i+1}\cdots p_j, p_1\cdots p_i)\top}([\boldsymbol{\mathcal{T}}]_j \otimes I_{p_{i+1}\cdots p_j}). \end{aligned}$$

Here, we define $e_k^{(ij)}$ as the kth canonical basis of $\mathbb{R}^{ij}$ and

$$A^{(i,j)} = \begin{bmatrix} e_1^{(ij)} & e_{i+1}^{(ij)} & \cdots & e_{i(j-1)+1}^{(ij)} \\ e_2^{(ij)} & e_{i+2}^{(ij)} & \cdots & e_{i(j-1)+2}^{(ij)} \\ \vdots & \vdots & \ddots & \vdots \\ e_i^{(ij)} & e_{2i}^{(ij)} & \cdots & e_{ij}^{(ij)} \end{bmatrix} \in \mathbb{R}^{(i^2 j) \times j}. \quad (5)$$

Lemmas 1 and 2 can be proved by checking each entry of the corresponding matricizations. In addition, the following lemma provides a representation of sequential reshaping tensor, in particular for $R_k^{(t)}$ and $\widetilde{R}_k^{(t)}$, the key intermediate outcomes in TTOI procedure.

Lemma 3 (Representation of Sequential Reshaping Tensor): Suppose $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{\otimes_{k=1}^d p_k}, M_i \in \mathbb{R}^{(r_{i-1}p_i) \times r_i}$ for $1 \le i \le d-1$, $B_i \in \mathbb{R}^{(p_i r_i) \times r_{i-1}}$ for $2 \le i \le d$, where $r_0 = r_d = 1$. Consider the following sequential multiplication:

1) Forward Sequential Multiplication: Let $S_1 = [\boldsymbol{\mathcal{T}}]_1$. For $k = 1, \ldots, d-1$, calculate

$$\begin{aligned} \widetilde{S}_k &= M_k^\top S_k \in \mathbb{R}^{r_k \times (p_{k+1}\cdots p_d)}, \\ S_{k+1} &= \text{Reshape}(\widetilde{S}_k, r_k p_{k+1}, p_{k+2}\cdots p_d) \quad \text{if } k < d-1. \end{aligned}$$

Then for any $1 \le k \le d-1$,

$$S_k = (I_{p_k} \otimes M_{\text{prod},k-1}^{(L)\top})[\boldsymbol{\mathcal{T}}]_k, \quad \widetilde{S}_k = M_{\text{prod},k}^{(L)\top}[\boldsymbol{\mathcal{T}}]_k. \quad (6)$$

Here, $I_{p_k} \otimes M_{\text{prod},k-1}^{(L)\top} = I_{p_1}$ if $k = 1$.

2) Backward Sequential Multiplication: Let $W_{d-1} = [\boldsymbol{\mathcal{T}}]_{d-1}$. For $k = d-1, \ldots, 1$, calculate

$$\begin{aligned} \widetilde{W}_k &= W_k B_{k+1} \in \mathbb{R}^{(p_1\cdots p_k) \times r_k}, \\ W_{k-1} &= \text{Reshape}(\widetilde{W}_k, p_1\cdots p_{k-1}, p_k r_k) \quad \text{if } k > 1. \end{aligned}$$

Then for any $1 \le k \le d-1$,

$$W_k = [\boldsymbol{\mathcal{T}}]_k (B_{\text{prod},k+2}^{(R)} \otimes I_{p_{k+1}}), \quad \widetilde{W}_k = [\boldsymbol{\mathcal{T}}]_k B_{\text{prod},k+1}^{(R)}.$$

Here, $B_{\text{prod},k+2}^{(R)} \otimes I_{p_{k+1}} = I_{p_d}$ if $k = d-1$.

In particular, $R_k^{(0)}, \widetilde{R}_k^{(0)}$ in Algorithm 1(a) and $R_k^{(t)}, \widetilde{R}_k^{(t)} (t \in \{2,4,6,\ldots\})$ in Algorithm 1(c) satisfy

$$R_k^{(t)} = \left(I_{p_k} \otimes (\widehat{U}^{(t)})^{(L)\top}_{\text{prod},k-1}\right)[\boldsymbol{\mathcal{Y}}]_k,$$
$$\widetilde{R}_k^{(t)} = (\widehat{U}^{(t)})^{(L)\top}_{\text{prod},k}[\boldsymbol{\mathcal{Y}}]_k, \quad \forall 1 \leq k \leq d-1. \tag{7}$$

The proof of Lemma 3 is provided in Section A-H.

B. Deterministic Upper Bounds for Estimation Error of TTOI

Now we are in position to analyze the performance of TTOI. The following Theorem 1 introduces an upper bound on estimation error of $\widehat{\boldsymbol{\mathcal{X}}}^{(2t+1)}$ (backward update) and $\widehat{\boldsymbol{\mathcal{X}}}^{(2t+2)}$ (forward update).

Theorem 1: Suppose we observe $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{Z}}$, where $\boldsymbol{\mathcal{X}}$ admits a TT decomposition as (1).

1) (A Deterministic Estimation Error Bound for Backward Updates): Let $\widetilde{U}_1^{(2t)} = U_1 \in \mathbb{R}^{p_1 \times r_1}$ be the left singular space of $[\boldsymbol{\mathcal{X}}]_1$. For $2 \leq k \leq d-1$, define $\widetilde{U}_k^{(2t)} \in \mathbb{R}^{p_k r_{k-1} \times r_k}$ as the left singular subspace of $\left(I_{p_k} \otimes (\widehat{U}^{(2t)})^{(L)\top}_{\text{prod},k-1}\right)[\boldsymbol{\mathcal{X}}]_k$. If for some constant $c_0 \in (0,1)$,

$$\left\|\sin\Theta\left(\widehat{U}_k^{(2t)}, \widetilde{U}_k^{(2t)}\right)\right\| \leq c_0, \quad \forall 1 \leq k \leq d-1, \tag{8}$$

then there exists a constant $C_d > 0$ that only depends on d such that the outcome of Algorithm 1(b) satisfies

$$\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(2t+1)} - \boldsymbol{\mathcal{X}}\right\|_{\text{F}}^2 \leq C_d \left(\sum_{k=1}^{d-1} A_k^{(2t+1)} + B^{(2t+1)}\right), \tag{9}$$

where

$$A_k^{(2t+1)} = \left\|(\widehat{U}^{(2t)})^{(L)\top}_{\text{prod},k}[\boldsymbol{\mathcal{Z}}]_k\left((\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},k+2} \otimes I_{p_{k+1}}\right)\right\|_{\text{F}}^2,$$
$$B^{(2t+1)} = \left\|[\boldsymbol{\mathcal{Z}}]_1(\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},2}\right\|_{\text{F}}^2.$$

Here, $(\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},k+2} \otimes I_{p_{k+1}} = I_{p_d}$ if $k = d-1$.

2) (A Deterministic Estimation Error Bound for Forward Updates): For $2 \leq k \leq d-1$, let $\widetilde{V}_k^{(2t+1)} \in \mathbb{R}^{(p_k r_k) \times r_{k-1}}$ be the right singular space of $[\boldsymbol{\mathcal{X}}]_{k-1}\left((\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},k+1} \otimes I_{p_k}\right)$ and let $\widetilde{V}_d^{(2t+1)} = V_d \in \mathbb{R}^{p_d \times r_{d-1}}$ be the right singular space of $[\boldsymbol{\mathcal{X}}]_{d-1}$. If for some constant $c_0 \in (0,1)$,

$$\left\|\sin\Theta\left(\widehat{V}_k^{(2t+1)}, \widetilde{V}_k^{(2t+1)}\right)\right\| \leq c_0, \quad \forall 2 \leq k \leq d,$$

then there exists a constant $C_d > 0$ that only depends on d such that the outcome of Algorithm 1(c) satisfies

$$\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(2t+2)} - \boldsymbol{\mathcal{X}}\right\|_{\text{F}}^2 \leq C_d \left(\sum_{k=1}^{d-1} A_k^{(2t+2)} + B^{(2t+2)}\right), \tag{10}$$

where

$$A_k^{(2t+2)} = \left\|\left(I_{p_k} \otimes (\widehat{U}^{(2t+2)})^{(L)\top}_{\text{prod},k-1}\right)[\boldsymbol{\mathcal{Z}}]_k(\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},k+1}\right\|_{\text{F}}^2,$$
$$B^{(2t+2)} = \left\|(\widehat{U}^{(2t+2)})^{(L)\top}_{\text{prod},d-1}[\boldsymbol{\mathcal{Z}}]_{d-1}\right\|_{\text{F}}^2.$$

Here, $I_{p_k} \otimes (\widehat{U}^{(2t+2)})^{(L)\top}_{\text{prod},k-1} = I_{p_1}$ if $k = 1$.

The proof of Theorem 1 is provided in Section A-A. Theorem 1 shows the estimation error $\|\widehat{\boldsymbol{\mathcal{X}}}^{(t+1)} - \boldsymbol{\mathcal{X}}\|_{\text{F}}^2$ can be bounded by the projected noise $\boldsymbol{\mathcal{Z}}$, i.e., $A_k^{(t+1)}$ and $B^{(t+1)}$, if the estimates in initialization ($t = 0$) or the previous iteration ($t \geq 1$), $\{\widehat{U}_k^{(t)}\}_{k=1}^{d-1}$ or $\{\widehat{V}_k^{(t)}\}_{k=2}^{d}$, are within constant distance to the true underlying subspaces. The developed upper bound can be significantly smaller than $C\|\boldsymbol{\mathcal{Z}}\|_{\text{F}}^2$, the classic upper bound induced from the approximation error (e.g., Theorem 2.2 in [17]), especially in the high-dimensional setting ($p \gg r$).

Remark 3 (Interpretation of Error Bounds in Theorem 1): Here, we provide some explanation for $A_k^{(2t+1)}$ and $B^{(2t+1)}$ in the error bound (9). By algebraic calculation, the TT-core estimation via backward update can be written as

$$\widehat{V}_{k+1}^{(2t+1)} = \text{SVD}^R\left\{(\widehat{U}^{(2t)})^{(L)\top}_{\text{prod},k}([\boldsymbol{\mathcal{X}}]_k + [\boldsymbol{\mathcal{Z}}]_k) \cdot \left((\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},k+2} \otimes I_{p_{k+1}}\right)\right\}$$

for any $1 \leq k \leq d-1$ and

$$\widehat{V}_1^{(2t+1)} = ([\boldsymbol{\mathcal{X}}]_1 + [\boldsymbol{\mathcal{Z}}]_1)(\widehat{V}^{(2t+1)})^{(R)}_{\text{prod},2}.$$

From the definition of $A_k^{(2t+1)}$, we have see $A_k^{(2t+1)}$ quantifies the error of the singular subspace estimate $\widehat{V}_{k+1}^{(2t+1)}$ and $B^{(2t+1)}$ quantifies the error of the projected residual $\widehat{V}_1^{(2t+1)}$. By symmetry, similar interpretation also applies to $A_k^{(2t+2)}$ and $B^{(2t+2)}$ for the error bound of forward update (10).

Remark 4 (Proof Sketch of Theorem 1): While the complete proof of Theorem 1 is provided in Section A-A, we provide a brief proof sketch here.

Without loss of generality, we focus on (9) for $t = 0$ while other cases follows similarly. For convenience, we simply let $\widehat{U}_i, \widehat{V}_i$ denote $\widehat{U}_i^{(0)}, \widehat{V}_i^{(1)}$, respectively. First, by Lemma 1, we can transform $[\widehat{X}^{(1)}]_1$, the outcome of backward update, to

$$[\widehat{X}^{(1)}]_1 = [\boldsymbol{\mathcal{Y}}]_1 P_{(\widehat{V}_d \otimes I_{p_2 \cdots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2}.$$

Then we can further bound the estimation error of $\widehat{\boldsymbol{\mathcal{X}}}^{(1)}$ as

$$\begin{aligned}
&\|\widehat{\boldsymbol{\mathcal{X}}}^{(1)} - \boldsymbol{\mathcal{X}}\|_{\text{F}}^2 \\
\leq & C\left\|[\boldsymbol{\mathcal{Z}}]_1(\widehat{V}_d \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2\right\|_{\text{F}}^2 \\
&+ C_d \sum_{k=2}^{d}\left\|[\boldsymbol{\mathcal{X}}]_1(\widehat{V}_d \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k}) \cdot (\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_{\text{F}}^2.
\end{aligned}$$

Next, based on Lemma 2 and (8), we can prove

$$\begin{aligned}
&\left\|[\boldsymbol{\mathcal{X}}]_1(\widehat{V}_d \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k}) \cdot (\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_{\text{F}} \\
= & \left\|[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_{\text{F}} \\
\leq & C_d\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1} \cdot (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_{\text{F}}.
\end{aligned}$$

Finally, we apply the perturbation projection error bound (Lemma 6) to prove that

$$\begin{aligned} &C_d\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes\widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1} \\ &\quad\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\mathrm{F}} \\ \le &C_d\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes\widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_{k-1} \\ &\quad\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\Big\|_{\mathrm{F}}. \end{aligned}$$

Theorem (1) is proved by combing all inequalities above.

Next, we establish a decomposition formula for the approximation error, i.e., the objective function in (3) $\left\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{X}}^{(t)}\right\|_{\mathrm{F}}^2$, and show that the approximation error is monotone decreasing through TTOI iterations.

Theorem 2 (Approximation error decays through iterations): We implement TTOI on $\boldsymbol{\mathcal{Y}}$. Let $\widehat{\boldsymbol{\mathcal{X}}}^{(t)}$ be the outcome after the tth iteration. For any $k\ge 1$, we have

(Approximation error decay)

$$\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}^2-\|\widehat{\boldsymbol{\mathcal{X}}}^{(t+1)}\|_{\mathrm{F}}^2\le\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}^2-\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}\|_{\mathrm{F}}^2, \tag{11}$$

(Approximation error decomposition)

$$\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(t+1)}\|_{\mathrm{F}}^2=\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}^2-\|\widehat{\boldsymbol{\mathcal{X}}}^{(t+1)}\|_{\mathrm{F}}^2. \tag{12}$$

See Section A-B for the proof of Theorem 2.

IV. TTOI for Tensor-Train Spiked Tensor Model

In this section, we further focus on a probabilistic setting, *spiked tensor model*, where the noise tensor $\boldsymbol{\mathcal{Z}}$ has independent, mean zero, and σ-sub-Gaussian entries (see definition in Section II-A). The spiked tensor model has been widely studied as a benchmark setting for tensor PCA/SVD and dimension reduction in recent literature in machine learning, information theory, statistics, and data science [48], [60]–[62], [70]. The central goal therein is to discover the underlying low-rank tensor $\boldsymbol{\mathcal{X}}$. Most of the existing works focused on tensors with Tucker or CP decomposition.

Under the spiked tensor model, we can verify that the initialization step of TTOI gives sufficiently good initial estimations with high probability that matches the required condition in Theorem 1.

Theorem 3 (Probabilistic Bound for Initial Estimates and Projected Noise): Suppose $\boldsymbol{\mathcal{X}}$ is TT-decomposable as (1) and $\boldsymbol{\mathcal{Z}}$ have independent zero mean and σ-sub-Gaussian random variables. Denote $p=\min\{p_1,\cdots,p_d\}$. If there exists a constant C_{gap} such that $\lambda_k=s_{r_k}([\boldsymbol{\mathcal{X}}]_k)\ge C_{gap}\left((\sum_{i=1}^d p_ir_{i-1}r_i)^{1/2}+(p_{k+1}\cdots p_d)^{1/2}\right)\sigma$ for $1\le k\le d-1$, then there exist some constants $C,c>0$ and $C_d>0$ that only depends on d, with probability at least $1-C\exp(-cp)$,

$$\max_{k=1,\ldots,d-1}\left\|\sin\Theta\left(\widehat{U}_k^{(0)},\widetilde{U}_k^{(0)}\right)\right\|\le\frac{1}{2}, \tag{13}$$

$$\max_{\substack{k=1,\ldots,d-1\\ t=2,4,6,\ldots}}\left\|\sin\Theta\left(\widehat{U}_k^{(t)},\widetilde{U}_k^{(t)}\right)\right\|\le\frac{1}{2},$$

$$\max_{\substack{k=2,\ldots,d\\ t=1,3,5,\ldots}}\left\|\sin\Theta\left(\widehat{V}_k^{(t)},\widetilde{V}_k^{(t)}\right)\right\|\le\frac{1}{2}, \tag{14}$$

and for all $t\ge 1$,

$$\max\{A_k^{(t)},B^{(t)}\}\le C_d\sigma^2\sum_{i=1}^d p_ir_ir_{r-1}. \tag{15}$$

Here, $\widetilde{U}_k^{(t)}$, $\widetilde{V}_k^{(t)}$, $A_k^{(t)}$ and $B^{(t)}$ are defined in Theorem 1.

The proof of Theorem 3 is provided in Section A-C. Based on Theorems 1 and 3, we can further prove:

Corollary 1 (Upper bound for estimation error): Suppose $\boldsymbol{\mathcal{X}}$ can be decomposed as (1), $\boldsymbol{\mathcal{Z}}_{i_1,\ldots,i_d}$ are independent zero mean and σ-sub-Gaussian random variables, $p=\min\{p_1,\cdots,p_d\}$. Suppose there exists a constant C_{gap} such that $\lambda_k=s_{r_k}([\boldsymbol{\mathcal{X}}]_k)\ge C_{gap}\left((\sum_{i=1}^d p_ir_{i-1}r_i)^{1/2}+(p_{k+1}\cdots p_d)^{1/2}\right)\sigma$ for $1\le k\le d-1$. Then with probability at least $1-Ce^{-cp}$, for all $t\ge 1$,

$$\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^2\le C_d\sigma^2\sum_{i=1}^d p_ir_ir_{i-1}. \tag{16}$$

The proof of Corollary 1 is provided in Section A-D.

Remark 5 (Interpretation of Corollary 1): Note that the TT-cores $G_1,\boldsymbol{\mathcal{G}}_i,G_d$ respectively have $p_1r_1,p_ir_ir_{i-1},p_dr_{d-1}$ free parameters, the upper bound (16) can be seen as the noise level σ^2 times the degrees of freedom of the low TT rank tensors.

Next, we develop a minimax lower bound for the low TT rank structure estimation. Consider the following general class of tensors with dimension $\boldsymbol{p}=(p_1,\ldots,p_d)$ and TT rank $\boldsymbol{r}=(r_1,\ldots,r_{d-1})$,

$$\begin{aligned}\mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda})=\big\{&\boldsymbol{\mathcal{X}}\in\mathbb{R}^{p_1\times\cdots\times p_d},\boldsymbol{\mathcal{X}}\text{ can be decomposed as (1)},\\ &s_{r_k}([\boldsymbol{\mathcal{X}}]_k)\ge\lambda_k,1\le k\le d-1\big\},\end{aligned} \tag{17}$$

and a class of distributions of σ-sub-Gaussian noise tensors

$$\begin{aligned}\mathcal{D}=\{D:\text{ if }\boldsymbol{\mathcal{Z}}\sim D,\text{ then }&\boldsymbol{\mathcal{Z}}_{i_1,\ldots,i_d}\text{ are indep. zero mean}\\ &\text{and }\sigma\text{ sub-Gaussian random variables}\}.\end{aligned} \tag{18}$$

Here, the constraints on the least singular value of $[\boldsymbol{\mathcal{X}}]_k$ and the σ-sub-Gaussian assumption correspond to the conditions required for upper bound in Theorem 3.

Theorem 4 (Lower Bound): Consider the order-d TT spiked tensor model (2) and distribution class $\mathcal{D}$ in (18). Assume $p=\min\{p_1,\ldots,p_d\}\ge C_0$ for some large constant C_0, $r_1\le p_1/2, r_i\le p_ir_{i-1}/2, r_{i-1}\le p_ir_i/2$ for $2\le i\le d-1$, $r_{d-1}\le p_d$, and $\lambda_i>0$. Also assume $r_1r_2\le p_1$ if $d=3$. Then there exists a constant $c_d>0$ that only depends on d such that

$$\inf_{\widehat{\boldsymbol{\mathcal{X}}}}\sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda}),D\in\mathcal{D}}\mathbb{E}_{\boldsymbol{\mathcal{Z}}\sim D}\left\|\widehat{\boldsymbol{\mathcal{X}}}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2\ge c_d\sigma^2\sum_{i=1}^d p_ir_ir_{i-1}. \tag{19}$$

See Section A-E for the proof of Theorem 4.

V. TTOI for Dimension Reduction and State Aggregation in High-Order Markov Chain

Since the introduction at the beginning of the 20th century, the Markov process has been ubiquitous in a variety of disciplines. In the literature, the first order Markov process, i.e., the future observation at $(t+1)$ is conditionally independent of those at times $1, \ldots, (t-1)$ given the immediate past observation at time t, has been commonly used and extensively studied. Moreover, the high-order Markov process often appear in many scenarios, where the future observation is affected by a longer history. For example, in the taxi travel trajectory, the future stop of a taxi not only depends on the current location but also the past path that reveals the direction this taxi is heading to [71]. The high-order Markov processes have also been applied to inter-personal relationship [72], financial econometrics [73], traffic flow [74], among many other applications.

We specifically consider an ergodic, time-invariant, and $(d-1)$st order Markov process on a finite state space $\{1, \ldots, p\}$. That is, the future state X_{t+d} depends on the current state X_{t+d-1} and the previous $(d-2)$ states $(X_{t+d-2}, \ldots, X_{t+1})$ jointly:

$$\mathbb{P}\left(X_{t+d}|X_1, \ldots, X_{t+d-1}\right) = \mathbb{P}\left(X_{t+d}|X_{t+1}, \ldots, X_{t+d-1}\right) = \boldsymbol{\mathcal{P}}_{[X_{t+1}, \ldots, X_{t+d}]}. \tag{20}$$

Our goal is to achieve a reliable estimation of the transition tensor $\boldsymbol{\mathcal{P}}$ and to predict the future state X_{t+d} based on an observable trajectory. Since the total number of free parameters in a $(d-1)$st order Markov transition tensor $\boldsymbol{\mathcal{P}}$ is $O(p^d)$ without further assumptions, it may be prohibitively difficult to infer $\boldsymbol{\mathcal{P}}$ in both statistics and computation even if p and d are only of moderate scale. Instead, a sufficient dimension reduction for high-order Markov processes is in demand.

To enable the statistical inference and dimension reduction for high-order Markov processes, a powerful tool, mixed transition distribution model (MTD), was introduced [72]. The MTD model assumes that the distribution of future state is a linear combination of the distributions associated with the $(d-1)$ immediate past states. The readers are also referred to [75] for a survey on mixed transition distribution model. The linear assumption, however, does not take into account the potential interactions of past states that commonly appear in practice. For example in the New York taxi trip data, the interaction among past locations of a taxi indicates its potential future direction.

On the other hand, there is a recent surge of development in dimension reduction and state aggregation for first order Markov chains. For example, [76] considered the Markov chain aggregation and the application to biology; [77] considered the rank-reduced Markov model and mode clustering; [45] considered Markov rank, aggregagability, and lumpability of Markov processes and proposed the dimension reduction and state aggregation methods through spectral decomposition with theoretical guarantees; [78] proposed clustering block model and proposed efficient algorithm to solve it; [79] introduced a convex and non-convex methods to estimate the rank-reduced low-rank Markov transition matrix.

Inspired by these work, we propose and study the *state aggregation model* for the discrete-time high-order Markov processes as follows.

Definition 1 ($(d-1)$st Order State Aggregatable Markov Process): Suppose there exist maps $G_1 : [p] \to \mathbb{R}^{r_1}$, $G_k : [p] \times \mathbb{R}^{r_{k-1}} \to \mathbb{R}^{r_k}$, $G_d : [p] \times \mathbb{R}^{r_{d-1}} \to \mathbb{R}$ such that $G_2, \ldots, G_d$ are linear: $G_k(X, \lambda_1 u + \lambda_2 v) = \lambda_1 G_k(X, u) + \lambda_2 G_k(X, v)$ for any vectors u, v, scalars $\lambda_1, \lambda_2 \in \mathbb{R}$. We say a Markov process $\{X_1, X_2, \ldots\}$ is $(d-1)$st order state aggregatable if for all $t \geq 0$, the transition can be sequentially generated as follows,

$$\begin{aligned}
&\widetilde{P}_1(X_{t+1}) = G_1(X_{t+1}) \in \mathbb{R}^{r_1}, \\
&\widetilde{P}_k(X_{t+1}, \ldots, X_{t+k}) = G_k(X_{t+k}, \widetilde{P}_{k-1}(X_{t+1}, \ldots, X_{t+k-1})) \\
&\qquad \in \mathbb{R}^{r_k}, \quad k = 2, \ldots, d-1, \\
&\mathbb{P}\left(X_{t+d}|X_1, \ldots, X_{t+d-1}\right) \\
&= \mathbb{P}\left(X_{t+d}|X_{t+1}, \ldots, X_{t+d-1}\right) \\
&= G_d(X_{t+d}, \widetilde{P}_{d-1}(X_{t+1}, \ldots, X_{t+d-1})).
\end{aligned}$$

In a $(d-1)$st order state aggregatable Markov process, the future state X_{t+d} relies on a sequential aggregation of the previous $d-1$ states $X_{t+1}, \ldots, X_{t+d-1}$ as follows: we first project X_{t+1} to a r_1-dimensional vector $\widetilde{P}_1(X_{t+1})$ via G_1, then project $\widetilde{P}_1(X_{t+1})$ jointly with X_{t+2} to a r_2-dimensional vector $\widetilde{P}_1(X_{t+1}, X_{t+2})$ via G_2. We repeat such the projection sequentially for $X_{t+3}, \ldots, X_{t+d}$ and yield the transition probability $\mathbb{P}\left(X_{t+d}|X_{t+1}, \ldots, X_{t+d-1}\right)$. Also, see Figure 4 for a pictorial illustration.

Based on the definition of the state aggregatable Markov chain, we can prove the corresponding probability transition tensor $\boldsymbol{\mathcal{P}}$ will have low TT rank.

Proposition 1: The transition tensor $\boldsymbol{\mathcal{P}}$ of the rank reduced high-order Markov model in Definition 1 has TT-rank no more than $(r_1, \ldots, r_{d-1})$. In other words, $\boldsymbol{\mathcal{P}}$ satisfies $\mathrm{rank}([\boldsymbol{\mathcal{P}}]_k) \leq r_k$.

The proof of Proposition 1 is provided in Section A-F.

Next, we focus on a *synchronous* or *generative setting*, which can be seen as a high-order generalization of the classic observation model for the analysis of Markov (decision/reward) processes (see [80] for an introduction), for the high-order Markov process. To be specific, for each sample index $k = 1, \ldots, n$ and previous states $(i_1, \ldots, i_{d-1}) \in [p]^{d-1}$, suppose we observe the next state $X(i_1, \ldots, i_{d-1}; k)$ drawn from the Markov transition tensor $\boldsymbol{\mathcal{P}}$. It is natural to estimate $\boldsymbol{\mathcal{P}}$ via the empirical transition tensor: for $i_1, \ldots, i_d \in \{1, \ldots, p\}^d$,

$$\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}_{i_1, \ldots, i_d} = \sum_{k=1}^{n} 1_{\{X(i_1, \ldots, i_{d-1}; k) = i_d\}} \Big/ n.$$

Then, $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$ is an unbiased estimator of $\boldsymbol{\mathcal{P}}$. However, if the entries of $\boldsymbol{\mathcal{P}}$ are approximately balanced, the mean squared error of $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$ satisfies

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}} - \boldsymbol{\mathcal{P}}\right\|_{\mathrm{F}}^2 = \sum_{i_1, \ldots, i_d} \mathrm{Var}\left(\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}_{i_1, \ldots, i_d}\right) \\
&= \sum_{i_1, \ldots, i_{d-1}} \sum_{i_d} \frac{\mathbb{P}\left(i_d|i_1, \ldots, i_{d-1}\right)\left(1 - \mathbb{P}\left(i_d|i_1, \ldots, i_{d-1}\right)\right)}{n} \\
&\asymp \frac{p^{d-1}}{n},
\end{aligned} \tag{21}$$

To obtain a more accurate estimator, we propose to first perform TTOI on $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$ to obtain $\widehat{\boldsymbol{\mathcal{P}}}^{(1)}$, then project each

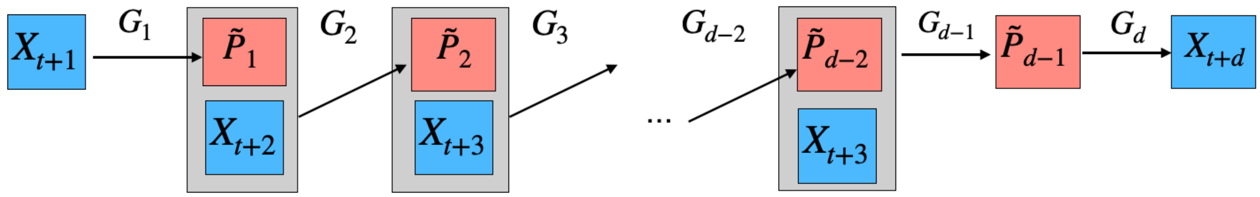


Fig. 4. A pictorial illustration of a $(d-1)$st order state aggregatable Markov chain.

row of $[\widehat{\mathcal{P}}^{(1)}]_{d-1}$, or equivalently, each mode-d fiber of $\widehat{\mathcal{P}}^{(1)}$, onto the simplex $S^{p-1} = \{x \in \mathbb{R}^p : \sum_{i=1}^p x_i = 1, x_i \geq 0 \text{ for all } 1 \leq i \leq p\}$ via probability simplex projection (see an implementation in [81]) and obtain $\widehat{\mathcal{P}}$.

We establish an upper bound on estimation error for the TTOI estimator $\widehat{\mathcal{P}}$.

Proposition 2: Consider the synchronous or generative model for a $(d-1)$st order state aggregatable Markov process described above. Suppose the initialization condition (8) in Theorem 1 holds. Then with probability at least $1-Ce^{-cp}$, the output of one-step TTOI followed by the probability simplex projection satisfies

$$\left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|_{\mathrm{F}}^2 \leq C\left(\max_{1\leq i\leq d-1} r_i\right)\sum_{i=1}^{d} p_i r_i r_{i-1}\Big/ n.$$

The proof of Proposition 2 is provided in Section A-G. Compared to the estimation error rate of $\widehat{\mathcal{P}}^{\mathrm{emp}}$ in (21), Proposition 2 shows TTOI achieves significantly reduced estimation error by exploiting the low TT rank structure of the high-order Markov process.

Remark 6: If the observations form one transition trajectory $\{X_0, \ldots, X_N\}$, we can work on the following empirical transition tensor:

$$\begin{aligned}
&\widehat{\mathcal{P}}^{\mathrm{emp}}_{i_1,\ldots,i_d} = \frac{\sum_{t=0}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-1}=i_d\}}}{\sum_{t=0}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}}},\\
&\text{ if } \sum_{t=1}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}} > 0;\\
&\widehat{\mathcal{P}}^{\mathrm{emp}}_{i_1,\ldots,i_d} = 1/p,\\
&\text{ if } \sum_{t=1}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}} = 0. \qquad (22)
\end{aligned}$$

Then $\widehat{\mathcal{P}}^{\mathrm{emp}}$ can be a nearly unbiased and strongly consistent estimator for $\mathcal{P}$. When the Markov process is $(d-1)$st order state aggregatable, we can apply TTOI to obtain a better estimate. As will be explored by numerical studies in Section VI-A, the TTOI estimator achieves favorable performance on the estimation of $\mathcal{P}$.

VI. Numerical Studies

In this section, we investigate the numerical performance of TTOI.

A. Simulation

In each simulation setting, we present the numerical results in both average estimation error (denoted by dots) and standard deviation (denoted by bars) based on 100 repetitions. We assume the true TT-ranks are known in the first three settings. Afterwards, we introduce a BIC-type data-driven scheme for TT-rank selection and present its numerical performance. All experiments are conducted by a quad-core 2.3 GHz Intel Core i5 processor.

We first consider the tensor-train spiked tensor model (2) discussed in Section IV. Specifically, we randomly generate $G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d$ with i.i.d. standard normal entries, and generate $\mathcal{Z}$ with i.i.d. $\mathcal{N}(0,\sigma^2)$ or Unif$(-b,b)$ entries. Let $p_1 = \cdots = p_d = p$, $r_1 = \cdots = r_{d-1} = r$, and consider four settings: (1) $p = 100, d = 3, r = 1$; (2) $p = 50, d = 4, r = 1$; (3) $p = 20, d = 5, r = 1$; (4) $p = 20, d = 5, r = 2$. For varying values of $\sigma \in [1, 19]$ and $b \in [3, 30]$, we evaluate the estimation error $\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_{\mathrm{F}}$ of the TT-SVD and TTOI estimators with 1 or 2 iterations, i.e., $t_{\max} = 0, 1, 2$. From the results summarized in Figure 5 (normal noise) and Figure 6 (uniform noise), we can see TTOI, even with one iteration, performs significantly better than TT-SVD, and the advantage becomes more significant as the noise level σ, b grows. This suggests that the proposed TTOI is effective for high-order tensor SVD compared to the classic TT-SVD, especially when the observations are corrupted by substantial noise. Table I summarizes the runtime of TT-SVD and TTOI, which suggests that the additional computational cost incurred by the backward and forward updates in TTOI is negligible compared to the runtime of the original TT-SVD.

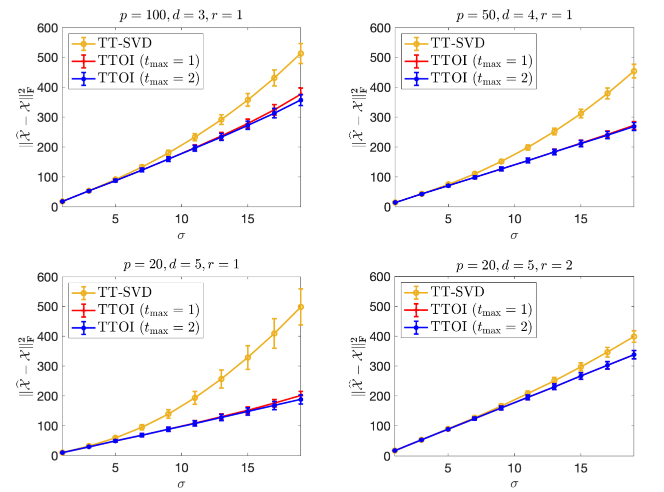


Fig. 5. Estimation error of TT-SVD and TTOI for high-order spiked tensor model. Here, $\mathcal{Z} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

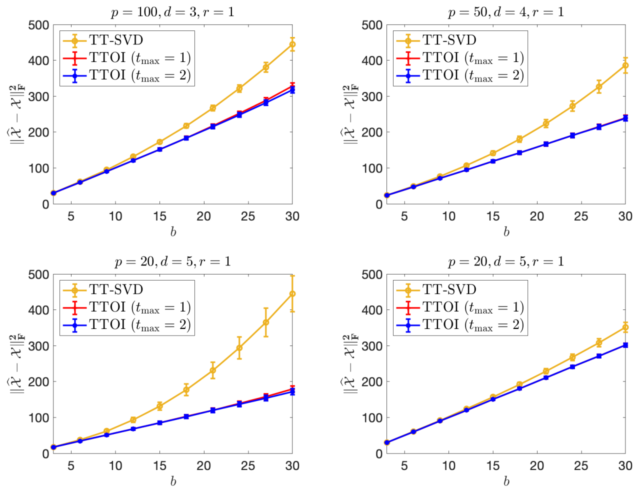


Fig. 6. Estimation error of TT-SVD and TTOI for high-order spiked tensor model. Here, $\mathcal{Z} \overset{\text{i.i.d.}}{\sim} \text{Unif}(-b, b)$.

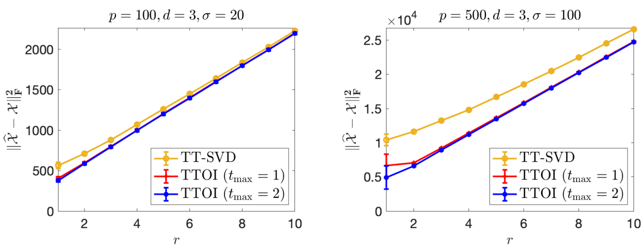


Fig. 7. Estimation error of TT-SVD and TTOI for high-order spiked tensor model with varying TT-ranks.

To understand the influence of TT-rank to the performance of the TT-SVD and TTOI estimators, we conduct numerical experiments under the spiked tensor model (2) with $r_1 = \cdots = r_{d-1} = r$ for various values of r. In particular, $G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d$ are still generated with i.i.d. standard normal entries, and $\mathcal{Z}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Letting $p_1 = \cdots = p_d = p$, we consider two settings: (1) $p = 100, d = 3, \sigma = 20$; (2) $p = 500, d = 3, \sigma = 100$. For $r = 1, \ldots, 10$, we evaluate the average estimation error $\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F$ of TT-SVD, TTOI with 1 iteration, and TTOI with 2 iterations (i.e., $t_{\max} = 0, 1, 2$), and present the results in Figure 7. Figure 7 suggests that the estimation errors increase as the rank increases, while TTOI with 1 or 2 iterations both performs better than TT-SVD. The improvement of TTOI over TT-SVD is more significant under larger p or smaller r. An intuitive explanation for this phenomenon is as follows: the key idea of TTOI is to utilize the previous updates to reduce the dimension of the sequential unfolding $[\mathcal{Y}]_k$ before performing singular value thresholding; such the dimension reduction is more significant for large p or small r.

Next, we demonstrate the performance of TTOI on transition tensor estimation for the high-order state-aggregatable Markov chains studied in Section V. We consider the $(d-1)$st order Markov chain on p states. To generate the transition tensor $\mathcal{P}$, we first draw $\widetilde{G}_1 \in \mathbb{R}^{p\times r}, \widetilde{\mathcal{G}}_2 \in \mathbb{R}^{r\times p\times r}, \ldots, \widetilde{G}_d \in$

TABLE I

RUNTIME (IN SECONDS) OF TT-SVD, TTOI WITH 1 ITERATION, AND TTOI WITH 2 ITERATIONS UNDER THE HIGH-ORDER SPIKED TENSOR MODEL WITH $\mathcal{Z} \overset{\text{I.I.D.}}{\sim} N(0, 400)$. THE MEAN RUNTIME OF 50 INDEPENDENT REPLICATES ARE PRESENTED AND THE STANDARD DEVIATIONS ARE LISTED IN PARENTHESES

(p, d, r)	TT-SVD	TTOI ($t_{\max} = 1$)	TTOI ($t_{\max} = 2$)
$(100, 3, 1)$	0.332 (0.071)	0.334 (0.071)	0.340 (0.074)
$(50, 4, 1)$	1.165 (0.173)	1.169 (0.172)	1.201 (0.171)
$(20, 5, 1)$	0.725 (0.093)	0.730 (0.092)	0.751 (0.095)
$(20, 5, 2)$	0.672 (0.100)	0.676 (0.101)	0.708 (0.103)

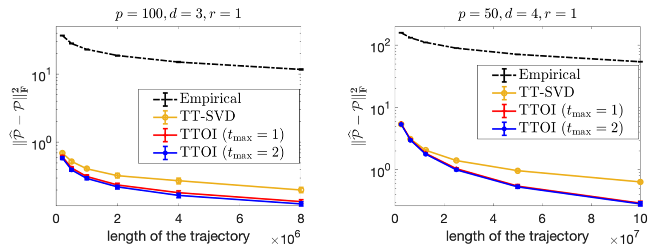


Fig. 8. Estimation error of the transition tensor versus length of the observable trajectory in high order state-aggregatable Markov chain estimation.

$\mathbb{R}^{r\times p}$ with i.i.d. standard normal entries, then normalize the rows of $\widetilde{G}_1, \widetilde{\mathcal{G}}_2, \ldots, \widetilde{G}_d$ in absolute values as

$$G_{1,[i,j]} = \frac{|\widetilde{G}_{1,[i,j]}|}{\sum_{j'} |\widetilde{G}_{1,[i,j']}|}, \quad \mathcal{G}_{k,[i_1,i_2,j]} = \frac{|\widetilde{\mathcal{G}}_{k,[i_1,i_2,j]}|}{\sum_{j'} |\widetilde{\mathcal{G}}_{k,[i_1,i_2,j']}|},$$
$$G_{d,[i,j]} = \frac{|\widetilde{G}_{d,[i,j]}|}{\sum_{j'} |\widetilde{G}_{d,[i,j']}|}.$$

By this means, $\mathcal{P} = [\![G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d]\!]$ satisfies $\mathcal{P}_{i_1,\ldots,i_d} \geq 0$, $\sum_{i_d=1}^{p} \mathcal{P}_{i_1,\ldots,i_d} = 1$ for any $(i_1, \ldots, i_{d-1})$, so $\mathcal{P}$ forms a Markov transition tensor. To generate the trajectory $\{X_1, \ldots, X_N\}$, we generate the initial $d-1$ states $X_1, \ldots, X_{d-1}$ i.i.d. uniformly from $[p]$, then generate $X_d, \ldots, X_N$ sequentially according to (20). To estimate $\mathcal{P}$, we construct the empirical probability tensor $\widehat{\mathcal{P}}^{\text{emp}}$ by (22), then apply TT-SVD and TTOI with input $\widehat{\mathcal{P}}^{\text{emp}}$ as detailed in Section V to obtain $\widehat{\mathcal{P}}$. We consider two numerical settings: (1) $p = 100, d = 3, r = 1$; (2) $p = 50, d = 4, r = 1$. We evaluate the estimation error $\|\widehat{\mathcal{P}}^{(i)} - \mathcal{P}\|_F$ for each setting and summarize the results to Figure 8. Again, TTOI exhibits clear advantage over the existing methods in all simulation settings.

1) Selection of TT-Ranks: The proposed TTOI algorithm requires specifying TT-ranks $r_1, \ldots, r_{d-1}$ as inputs and the appropriate choices of $r_1, \ldots, r_{d-1}$ are crucial in practice. We propose a data-driven scheme to select the TT-ranks: we choose $r_1, \ldots, r_{d-1} \geq 1$ such that the following Bayesian information criterion (BIC) under the spiked tensor model is minimized:

$$\begin{aligned} &\text{BIC}(r_1, \ldots, r_{d-1}) \\ &:= \prod_{k=1}^{d} p_k \log \|\mathcal{Y} - \widehat{\mathcal{X}}(r_1, \ldots, r_{d-1})\|_F^2 \end{aligned}$$

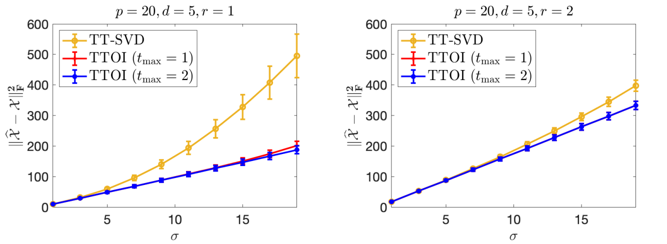


Fig. 9. Average estimation error of TT-SVD and TTOI for high-order spiked tensor model with BIC-tuned ranks.

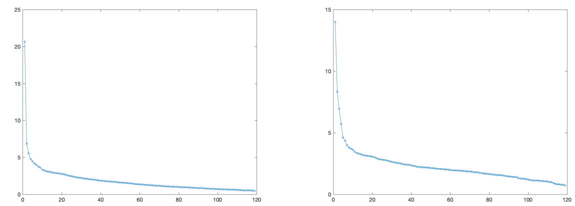


Fig. 10. Singular values of sequential unfolding matrices $[\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}]_1$ (left panel) and $[\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}]_2$ (right panel).

$$+\left(p_1 r_1 + \sum_{k=2}^{d-1} p_k r_{k-1} r_k + p_d r_{d-1}\right)\left(\sum_{k=1}^{d} \log p_k\right). \quad (23)$$

Here, $\widehat{\boldsymbol{\mathcal{X}}}(r_1, \ldots, r_{d-1})$ is the output of TTOI (Algorithm 1) with the input TT-ranks $r_1, \ldots, r_{d-1}$. This BIC-type criterion was also adopted in prior works on tensor clustering [82].

Then we conduct numerical experiments under the same setting as the bottom two plots in Figure 5 on the spiked tensor model with Gaussian noise. Figure 9 summarizes the estimation errors of TT-SVD and TTOI with 1 and 2 iterations, respectively, with the ranks selected based on the proposed BIC criterion (23). Comparing Figure 9 to the bottom two plots in Figure 5, we can see the proposed criterion can select the true ranks accurately and the performance of both TT-SVD and TTOI with tuned ranks is very similar to the one by inputting the true ranks.

B. Real Data Experiments

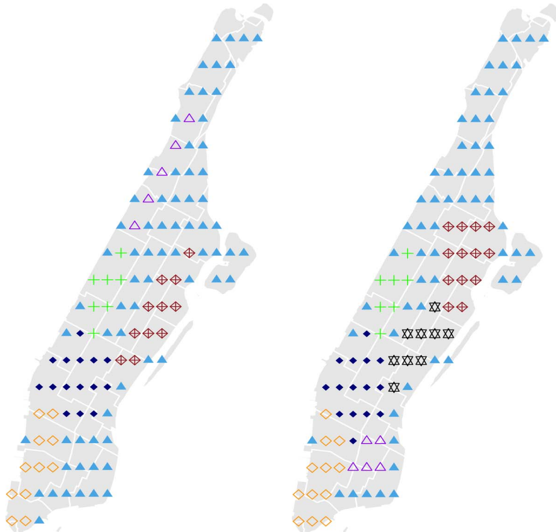
We apply the proposed method to investigate the Manhattan taxi data[1]. This dataset contains the New York City taxi trip records from 14,144 drivers in 2013. We treat each travel record as a transition among different locations at New York City, then the overall dataset can be organized as a collection of fragmented sample trajectories of a Markov chain on New York City traffic. Some recent analysis on such data can be seen at, e.g., [45], [71], [83].

Due to the high-dimensional spatiotemporal nature of the dataset, a sufficient dimension reduction or state aggregation is often a crucial first step to study a metropolitan-wide traffic pattern. To this end, we apply the high-order Markov model as described in Section V. Specifically, we discretize the Manhattan region into a grid of $p = 119$ states that forms a state space. Then, we collect all travel records in Manhattan of each driver from the dataset, sort them by time, and form into Markovian transition trajectories. In particular, each travel record is treated as a transition from the pickup to the drop-off location. If the drop-off location i of the previous trip is different from the pickup location j of the next trip by the same driver, we also form a transition from states i to j. Based on the trajectories, we can construct a high-order Markov chain with an order d empirical transition probability tensor $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}} \in \mathbb{R}^{\otimes_{k=1}^{d} p}$ as described in Section V. Assuming the true probability tensor is state aggregatable (Definition 1), we apply one-step TTOI proposed in Section V and obtain $\widehat{\boldsymbol{\mathcal{P}}}$. It is noteworthy if $d = 2$, the described procedure of $\widehat{\boldsymbol{\mathcal{P}}}$ is equivalent to the classic matrix spectral decomposition in the literature. Figure 10 plots the singular values of the sequential unfolding matrices of $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$ for $d = 3$, which clearly demonstrates the low-TT-rankness of the probability transition tensor $\boldsymbol{\mathcal{P}}$. In the following experiments, we focus on the order-2 Markov model and analyze all consecutive two transitions: $i \to j \to k$, corresponding to the $d = 3$ case.

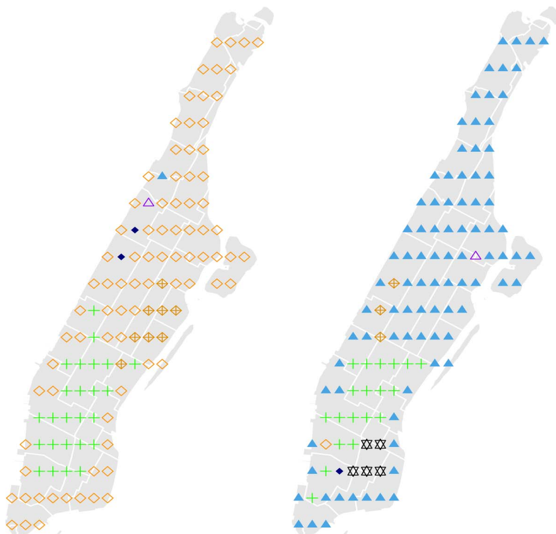
[1]2013 Trip Data, available at https://chriswhong.com/open-data/foil_nyc_taxi/

Inspired by the classic methods of matrix spectral decomposition, we aggregate all location states in Manhattan into a few clusters via both $\widehat{\boldsymbol{\mathcal{P}}}$ and $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$. Specifically, we calculate $\widehat{G}_d^\top$, i.e., the last TT-core of $\widehat{\boldsymbol{\mathcal{P}}}$, and $[\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}]_{d-1}$, i.e., the matricization of $\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}$ whose columns correspond to the last mode. Then we perform k-means on all columns of $\widehat{G}_d^\top$ and $[\widehat{\boldsymbol{\mathcal{P}}}^{\mathrm{emp}}]_{d-1}$, record the cluster index, associate the index to each location state, and plot the results in Figure 11 (Panels (a)(b) are for TTOI and Panels (c)(d) are for empirical estimate). From Figure 11 (a)(b), we can clearly identify four regions: (i) lower Manhattan (orange), (ii) midtown (dark blue), (iii) upper west side (green), and (iv) upper east side (brown or black). In contrast, direct clustering on $\boldsymbol{\mathcal{P}}^{\mathrm{emp}}$ yields less interpretable results as the majority points go to one cluster. It is also worth noting even the location information is not provided to this experiment, the resulting clusters in Figures 11 (a)(b) show good spatial proximity between locations. This illustrates the effectiveness of TTOI in dimension reduction and state-aggregation for high-order Markov processes.

Next, we illustrate the high-order nature of the city-wide taxi trip through the following experiment. For each initial state $i \in [p]$, we apply k-means to cluster the column span of $\widehat{\boldsymbol{\mathcal{P}}}_{[i,:,:]}$, where $\widehat{\boldsymbol{\mathcal{P}}}$ is the outcome of TTOI. We present the results in Figure 12, where the red triangles denote the given first state i and $r = k = 7$. If the city-wide taxi trips do not have significant high-order effects, $\widehat{\boldsymbol{\mathcal{P}}}$ should be reducible to a first order Markov process and $\widehat{\boldsymbol{\mathcal{P}}}_{[i,:,:]}$ should have similar values for different i. However, as we can see from Figure 12 that the clustering results highly depends on the first state i, the high-order effects exist in the city-wide taxi trip Markov process. In addition, the states in different directions of i are often clustered to different regions, which shows that the taxi drivers may tend to move to the same direction in consecutive trips, which yields the high-order effects in the driving trajectories.



(a) $\widehat{G}_d$, $r = 6, k = 6$ (b) $\widehat{G}_d$, $r = 7, k = 7$



(c) $[\widehat{\boldsymbol{\mathcal{P}}}^{\text{emp}}]_{d \quad 1} \quad k = 6$ (d) $[\widehat{\boldsymbol{\mathcal{P}}}^{\text{emp}}]_{d \quad 1} \quad k = 7$

Fig. 11. State aggregation based on TTOI and empirical estimate.

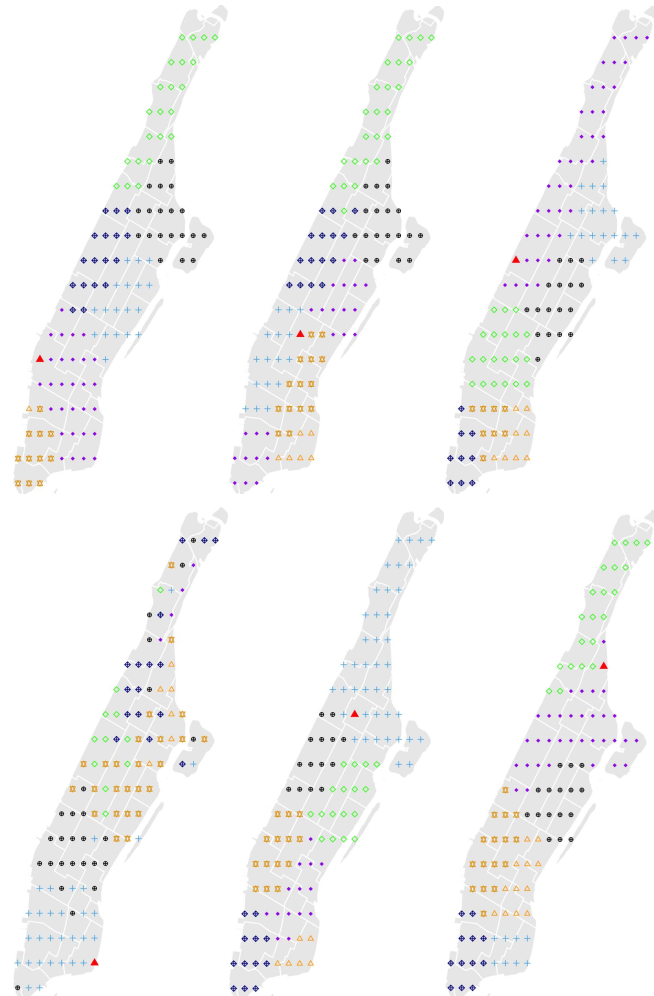


Fig. 12. Based on second order Markov model, state aggregation results are different with different initial state (the red triangle denotes the initial state i in each subfigure).

VII. Discussions and Additional Applications

In this paper, we propose a general framework for high-order SVD. We introduce a novel procedure, tensor-train orthogonal iteration (TTOI), that efficiently estimates the low tensor train rank structure from the high-order tensor observation. TTOI has significant advantages over the classic ones in the literature. We establish a general deterministic error bound for TTOI with the support of several new representation lemmas for tensor matricizations. Under the commonly studied spiked tensor model, we establish an upper bound for TTOI and a matching information-theoretic lower bound. We also illustrate the merits of TTOI through simulation studies and a real data example in New York City taxi trips.

In addition to the high-order Markov processes, the proposed TTOI can also be applied to the *Markov random field (MRF)* estimation. We give a brief description of MRF below. Consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, d\}$ is a set of vertices and $E \subseteq V \times V$ is a collection of edges. Each vertex $i \in V$ is associated with a random variable X_i, taking values in $\{s_1, \ldots, s_p\}$. In an MRF model, the distribution of $\{X_1, \ldots, X_d\}$ can be factorized as

$$\mathbb{P}(X_1, \ldots, X_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C),$$

where $\mathcal{C}$ is a collection of subgraphs of G and $X_C = (X_v, v \in C)$ denotes the random vector corresponding to vertices in C. The joint probability function $\mathbb{P}(\cdot)$ can be written as a tensor $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{\otimes_{k=1}^d p}$, where $\boldsymbol{\mathcal{P}}_{i_1,\ldots,i_d} = \mathbb{P}(X_1 = s_{i_1}, \ldots, X_d = s_{i_d})$. The MRFs have a wide range of applications, including image analysis [84], [85], genomic study [86], and natural language processing [87]. The readers are referred to, e.g., [88] for an introduction to MRFs.

A central problem of MRF is how to estimate the population density $\boldsymbol{\mathcal{P}}$ based on a limited number of samples $\{X_1^{(i)}, \ldots, X_d^{(i)}\}_{i=1}^n$. It is straightforward to estimate $\boldsymbol{\mathcal{P}}$ via the empirical probability tensor $\widehat{\boldsymbol{\mathcal{P}}}^{\text{emp}}$:

$$\widehat{\boldsymbol{\mathcal{P}}}^{\text{emp}}_{i_1,\ldots,i_d} = \sum_{i=1}^n \prod_{k=1}^d 1(X_k^{(i)} = s_{i_k}) \Big/ n.$$

We can show that $\widehat{\mathcal{P}}^{\text{emp}}$ is unbiased for $\mathcal{P}$. Recently, [16] pointed out that $\mathcal{P}$ is often approximately low tensor-train rank in practice. To further exploit such the structure, we can conduct TTOI on $\widehat{\mathcal{P}}^{\text{emp}}$. Under regularity conditions, it can be shown that the entries of $\mathcal{Z}$ are bounded and weakly independent, then Corollary 1 suggests the following estimation error rate of the TTOI estimator: $\|\widehat{\mathcal{P}} - \mathcal{P}\|_F^2 \leq C\sum_{i=1}^d r_i r_{i-1}/(np^{2d-1})$, which can be significantly smaller than the estimation error of original empirical estimator $\widehat{\mathcal{P}}^{\text{emp}}$.

Moreover, the proposed framework can be also applied to *high-order Markov decision process (high-order MDP)*. MDP has been commonly used as a baseline in control theory and reinforcement learning [89]–[92]. Despite the wide applications of MDPs, most of the existing work focus on the first-order Markov processes. However, the high-order effects often appear, i.e., the transition probability at the current time depends not only on current, but also the past $(d-1)$ states and actions. See Figure 13 for an example. Since the number of free parameters in such MDPs can be huge, a sufficient dimension reduction for the state and action space can be a crucial first step. Similarly to the example of high-order Markov process in Section V, the TTOI can be applied to achieve better dimension reduction and state aggregation for the high-order Markov decision processes.

Appendix A
Proofs

We collect all technical proofs of this paper in this section.

A. Proof of Theorem 1

For convenience, let $\widehat{U}_i$, $\widehat{V}_i$, R_i and $\widetilde{R}_i$ denote $\widehat{U}_i^{(0)}$, $\widehat{V}_i^{(1)}$, $R_i^{(0)}$ and $\widetilde{R}_i^{(0)}$, respectively. By Lemma 1 and

$$\begin{aligned} &I_{p_2\cdots p_d} - P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2} \\ =&P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_{2\perp}} \\ &+ P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_4\otimes I_{p_2p_3})(\widehat{V}_{3\perp}\otimes I_{p_2})} \\ &+\cdots+ P_{\widehat{V}_{d\perp}\otimes I_{p_2\cdots p_{d-1}}}, \end{aligned}$$

we have

$$\begin{aligned} &\left\|\widehat{\mathcal{X}}^{(1)} - \mathcal{X}\right\|_{\mathrm{F}}^2 \\ =&\Big\| \Big[[\mathcal{Y}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2\Big] \\ &\cdot \widehat{V}_2^\top(\widehat{V}_3^\top\otimes I_{p_2})\cdots(\widehat{V}_d^\top\otimes I_{p_2\cdots p_{d-1}}) - [\mathcal{X}]_1\Big\|_{\mathrm{F}}^2 \\ =&\Big\|[\mathcal{Z}]_1 P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2} \\ &+[\mathcal{X}]_1 P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2} - [\mathcal{X}]_1\Big\|_{\mathrm{F}}^2 \\ \leq& C\Big(\Big\|[\mathcal{Z}]_1 P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2}\Big\|_{\mathrm{F}}^2 \\ &+\Big\|[\mathcal{X}]_1 P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_{2\perp}}\Big\|_{\mathrm{F}}^2 \\ &+\Big\|[\mathcal{X}]_1 P_{(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_4\otimes I_{p_2p_3})(\widehat{V}_{3\perp}\otimes I_{p_2})}\Big\|_{\mathrm{F}}^2 \\ &+\cdots+\Big\|[\mathcal{X}]_1 P_{\widehat{V}_{d\perp}\otimes I_{p_2\cdots p_{d-1}}}\Big\|_{\mathrm{F}}^2\Big) \\ \leq& C\Big(\Big\|[\mathcal{Z}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_2\Big\|_{\mathrm{F}}^2 \\ &+\Big\|[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_{2\perp}\Big\|_{\mathrm{F}}^2 \\ &+\Big\|[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_4\otimes I_{p_2p_3})(\widehat{V}_{3\perp}\otimes I_{p_2})\Big\|_{\mathrm{F}}^2 \\ &+\cdots+\Big\|[\mathcal{X}]_1(\widehat{V}_{d\perp}\otimes I_{p_2\cdots p_{d-1}})\Big\|_{\mathrm{F}}^2\Big). \end{aligned} \tag{24}$$

To prove (9), we only need to show that for all $2\leq k\leq d$,

$$\begin{aligned} &\Big\|[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_2\cdots p_k}) \\ &\cdot(\widehat{V}_{k\perp}\otimes I_{p_2\cdots p_{k-1}})\Big\|_{\mathrm{F}} \\ \leq& C\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes\widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes\widehat{U}_1^\top)[\mathcal{Z}]_{k-1} \\ &\cdot(\widehat{V}_d\otimes I_{p_k\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\Big\|_{\mathrm{F}}, \end{aligned} \tag{25}$$

where

$$\begin{aligned} &[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_2\cdots p_k})(\widehat{V}_{k\perp}\otimes I_{p_2\cdots p_{k-1}}) \\ &=[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3\otimes I_{p_2})\widehat{V}_{2\perp} \end{aligned}$$

if $k=2$ and

$$\begin{aligned} &[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_2\cdots p_k})(\widehat{V}_{k\perp}\otimes I_{p_2\cdots p_{k-1}}) \\ &=[\mathcal{X}]_1(\widehat{V}_{d\perp}\otimes I_{p_2\cdots p_{d-1}}) \end{aligned}$$

if $k=d$.

By Lemma 2, we have

$$\begin{aligned} &\Big\|[\mathcal{X}]_1(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_2\cdots p_k}) \\ &\cdot(\widehat{V}_{k\perp}\otimes I_{p_2\cdots p_{k-1}})\Big\|_{\mathrm{F}} \\ =&\Big\|[A^{(p_2\cdots p_{k-1},p_1)}]^\top([\mathcal{X}]_{k-1}\otimes I_{p_2\cdots p_{k-1}})(\widehat{V}_d\otimes I_{p_2\cdots p_{d-1}})\cdots \\ &(\widehat{V}_{k+1}\otimes I_{p_2\cdots p_k})(\widehat{V}_{k\perp}\otimes I_{p_2\cdots p_{k-1}})\Big\|_{\mathrm{F}} \\ =&\Big\|[A^{(p_2\cdots p_{k-1},p_1)}]^\top\cdot\Big(\big([\mathcal{X}]_{k-1}(\widehat{V}_d\otimes I_{p_k\cdots p_{d-1}})\cdots \\ &(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\big)\otimes I_{p_2\cdots p_{k-1}}\Big)\Big\|_{\mathrm{F}} \\ =&\Big\|[\mathcal{X}]_{k-1}(\widehat{V}_d\otimes I_{p_k\cdots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\mathrm{F}}. \end{aligned} \tag{26}$$

The third equation holds since the realignment doesn't change the Frobenious norm.

Moreover, recall that $U_1\in\mathbb{R}^{p_1\times r_1}$ is the left singular space of $[\mathcal{X}]_1$, and $\widetilde{U}_j\in\mathbb{R}^{p_jr_{j-1}\times r_j}$ is the left singular space of $(I_{p_j}\otimes\widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j}\otimes\widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_j}\otimes\widehat{U}_1^\top)[\mathcal{X}]_j$ for $2\leq j\leq d-1$, by Lemma 2, for any $2\leq k\leq d-1$,

$$\begin{aligned} [\mathcal{X}]_k =&(I_{p_2\cdots p_k}\otimes[\mathcal{X}]_1)A^{(p_2\cdots p_k,p_{k+1}\cdots p_d)} \\ =&(I_{p_2\cdots p_k}\otimes P_{U_1}[\mathcal{X}]_1)A^{(p_2\cdots p_k,p_{k+1}\cdots p_d)} \\ =&(I_{p_2\cdots p_k}\otimes P_{U_1})(I_{p_2\cdots p_k}\otimes[\mathcal{X}]_1)A^{(p_2\cdots p_k,p_{k+1}\cdots p_d)} \\ =&(I_{p_2\cdots p_k}\otimes P_{U_1})[\mathcal{X}]_k, \end{aligned} \tag{27}$$

and for any $2\leq j<k$,

$$\begin{aligned} &(I_{p_j\cdots p_k}\otimes\widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k}\otimes\widehat{U}_{j-2}^\top)\cdots \\ &(I_{p_2\cdots p_k}\otimes\widehat{U}_1^\top)[\mathcal{X}]_k \end{aligned}$$

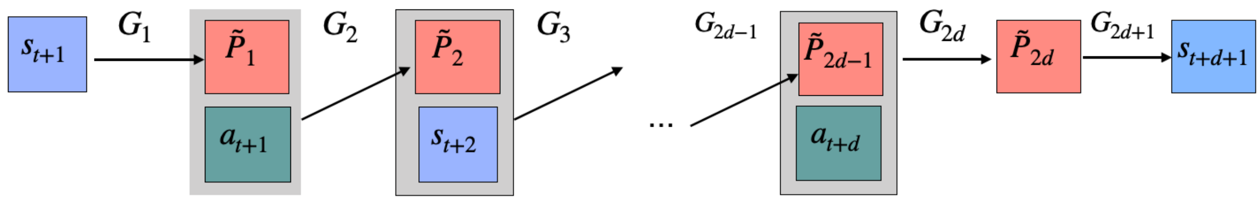


Fig. 13. Illustration of a high-order state aggregatable Markov decision process.

$$
\begin{aligned}
=&(I_{p_j\cdots p_k}\otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k}\otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_k}\otimes \widehat{U}_1^\top)\\
&\cdot(I_{p_{j+1}\cdots p_k}\otimes[\boldsymbol{\mathcal{X}}]_j)A^{(p_{j+1}\cdots p_k,p_{k+1}\cdots p_d)}\\
=&\Big(I_{p_{j+1}\cdots p_k}\otimes\big[(I_{p_j}\otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j}\otimes \widehat{U}_{j-2}^\top)\cdots\\
&\qquad(I_{p_2\cdots p_j}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\big]\Big)\\
&\cdot A^{(p_{j+1}\cdots p_k,p_{k+1}\cdots p_d)}\\
=&\Big(I_{p_{j+1}\cdots p_k}\otimes\big[P_{\widetilde{U}_j}(I_{p_j}\otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j}\otimes \widehat{U}_{j-2}^\top)\cdots\\
&\qquad(I_{p_2\cdots p_j}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\big]\Big)\\
&\cdot A^{(p_{j+1}\cdots p_k,p_{k+1}\cdots p_d)}\\
=&(I_{p_{j+1}\cdots p_k}\otimes P_{\widetilde{U}_j})(I_{p_j\cdots p_k}\otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k}\otimes \widehat{U}_{j-2}^\top)\cdots\\
&(I_{p_2\cdots p_k}\otimes \widehat{U}_1^\top)(I_{p_{j+1}\cdots p_k}\otimes[\boldsymbol{\mathcal{X}}]_j)A^{(p_{j+1}\cdots p_k,p_{k+1}\cdots p_d)}\\
=&(I_{p_{j+1}\cdots p_k}\otimes P_{\widetilde{U}_j})(I_{p_j\cdots p_k}\otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k}\otimes \widehat{U}_{j-2}^\top)\cdots\\
&(I_{p_2\cdots p_k}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_k,
\end{aligned}\tag{28}
$$

where $A^{(i,j)}$ is defined in (5) for any $i,j>0$.
Therefore, by (27),

$$
\begin{aligned}
&\Big\|[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
=&\Big\|(I_{p_2\cdots p_{k-1}}\otimes P_{U_1})[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
=&\Big\|(I_{p_2\cdots p_{k-1}}\otimes U_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
\le&\Big\|(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)(I_{p_2\cdots p_{k-1}}\otimes U_1)(I_{p_2\cdots p_{k-1}}\otimes U_1^\top)\\
&\cdot[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
&\cdot s_{\min}^{-1}\Big((I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)(I_{p_2\cdots p_{k-1}}\otimes U_1)\Big)\\
=&\Big\|(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\cdot s_{\min}^{-1}(\widehat{U}_1^\top U_1).
\end{aligned}\tag{29}
$$

The inequality holds since $\|B\|_{\rm F}\le\|AB\|_{\rm F}\cdot s_{\min}^{-1}(A)$ for any invertible matrix $A\in\mathbb{R}^{m_1\times m_1}$ and $B\in\mathbb{R}^{m_1\times m_2}$; in the last step, we used $(I_{p_2\cdots p_{k-1}}\otimes U_1)(I_{p_2\cdots p_{k-1}}\otimes U_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}=(I_{p_2\cdots p_{k-1}}\otimes P_{U_1})[\boldsymbol{\mathcal{X}}]_{k-1}=[\boldsymbol{\mathcal{X}}]_{k-1}$. Similarly to (29), by (28), for $1\le j\le k-2$,

$$
\begin{aligned}
&\Big\|(I_{p_{j+1}\cdots p_{k-1}}\otimes \widehat{U}_j^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
=&\Big\|(I_{p_{j+2}\cdots p_{k-1}}\otimes P_{\widetilde{U}_{j+1}})(I_{p_{j+1}\cdots p_{k-1}}\otimes \widehat{U}_j^\top)\cdots\\
&(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
=&\Big\|(I_{p_{j+2}\cdots p_{k-1}}\otimes \widetilde{U}_{j+1}^\top)(I_{p_{j+1}\cdots p_{k-1}}\otimes \widehat{U}_j^\top)\cdots\\
&(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)\cdot[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
\le&\Big\|(I_{p_{j+2}\cdots p_{k-1}}\otimes \widehat{U}_{j+1}^\top)(I_{p_{j+1}\cdots p_{k-1}}\otimes \widehat{U}_j^\top)\\
&\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\cdot s_{\min}^{-1}(\widehat{U}_{j+1}^\top\widetilde{U}_{j+1}).
\end{aligned}\tag{30}
$$

By (29) and (30),

$$
\begin{aligned}
&\Big\|[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
\le&\Big\|(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots\\
&(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\cdot s_{\min}^{-1}(\widehat{U}_1^\top U_1)\\
\le&\Big\|(I_{p_3\cdots p_{k-1}}\otimes \widehat{U}_2^\top)(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\cdot s_{\min}^{-1}(U_1^\top\widehat{U}_1)\\
&\cdot s_{\min}^{-1}(\widetilde{U}_2^\top\widehat{U}_2)\\
\le&\dots\\
\le&\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
&\cdot s_{\min}^{-1}(U_1^\top\widehat{U}_1)s_{\min}^{-1}(\widetilde{U}_2^\top\widehat{U}_2)\cdots s_{\min}^{-1}(\widetilde{U}_{k-1}^\top\widehat{U}_{k-1})\\
\le&\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}\\
&\cdot\left(\frac{1}{\sqrt{1-c_0^2}}\right)^{k-1}\\
\le& C\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\rm F}.
\end{aligned}\tag{31}
$$

By the definition of $\widehat{V}_k\in\mathbb{R}^{(p_kr_k)\times r_{k-1}}$ and Lemma 3, we know that $\widehat{V}_k$ is the right singular space of

$$
\begin{aligned}
&\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Y}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})
\end{aligned}
$$

$$
\begin{aligned}
=&\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\\
&+\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_{k-1}\\
&\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k}),
\end{aligned}
$$

Lemma 6 shows that

$$
\begin{aligned}
&\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{k-1}\\
&\quad\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_{\mathrm{F}}\\
\le&2\Big\|\widehat{U}_{k-1}^\top(I_{p_{k-1}}\otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}}\otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_{k-1}\\
&\quad\cdot(\widehat{V}_d\otimes I_{p_k\ldots p_{d-1}})\cdots(\widehat{V}_{k+1}\otimes I_{p_k})\Big\|_{\mathrm{F}}. \qquad (32)
\end{aligned}
$$

Combine (26), (31) and (32) together, we know that (25) holds for all $2\le k\le d$, which has finished the proof of Theorem 1.

B. Proof of Theorem 2

For $i\ge 1$, by the definition of $\boldsymbol{\mathcal{X}}^{(2i)}$ and Lemma 1, we have

$$
\begin{aligned}
&\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i)}\right\|_{\mathrm{F}}^2\\
=&\Big\|\Big(I_{p_1\cdots p_{d-1}}-P_{(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)})\cdots(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)})\widehat{U}_{d-1}^{(2i)}}\Big)\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_{d-1}\|_{\mathrm{F}}^2\\
&-\Big\|P_{(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)})\cdots(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)})\widehat{U}_{d-1}^{(2i)}}[\boldsymbol{\mathcal{Y}}]_{d-1}\Big\|_{\mathrm{F}}^2\\
=&\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}^2-\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(2i)}\right\|_{\mathrm{F}}^2.
\end{aligned}
$$

Similarly, we have

$$
\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i-1)}\right\|_{\mathrm{F}}^2=\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}^2-\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(2i-1)}\right\|_{\mathrm{F}}^2.
$$

In addition, we have

$$
\begin{aligned}
&\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i)}\right\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_{d-1}\|_{\mathrm{F}}^2\\
&-\Big\|P_{(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)})\cdots(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)})\widehat{U}_{d-1}^{(2i)}}[\boldsymbol{\mathcal{Y}}]_{d-1}\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_{d-1}\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_d^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_{d\perp}^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
\le&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_d^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2-\Big\|(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots\\
&\qquad(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})\cdot[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_d^{(2i-1)}\Big\|_{\mathrm{F}}^2.
\end{aligned}
$$

The last equation holds since $\widehat{U}_{d-1}^{(2i)}$ is the left singular space of $(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_d^{(2i-1)}$.

For any $B\in\mathbb{R}^{n\times r}$ and $1\le l\le r$, we can check that the l-th columns of $A^{(m,n)}B$ and $(I_m\otimes B\otimes I_m)A^{(m,r)}$ are equal:

$$
\begin{aligned}
(A^{(m,n)}B)_{[:,l]}=&\sum_{j=1}^n B_{j,l}\sum_{k=1}^m e_{(k-1)mn+(j-1)m+k}^{(m^2n)}\\
=&((I_m\otimes B\otimes I_m)A^{(m,r)})_{[:,l]}
\end{aligned}
$$

where $e_{(k-1)mn+(j-1)m+k}^{(m^2n)}$ is the $((k-1)mn+(j-1)m+k)$-th canonical basis of $\mathbb{R}^{m^2n}$ and $A^{(i,j)}$ is defined in (5). Therefore,

$$
A^{(m,n)}B=(I_m\otimes B\otimes I_m)A^{(m,r)}.
$$

By the last equation and Lemma 2, we have

$$
\begin{aligned}
&(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots\\
&(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})[\boldsymbol{\mathcal{Y}}]_{d-1}\widehat{V}_d^{(2i-1)}\\
=&(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots\\
&(I_{p_2\cdots p_{d-1}}\otimes\widehat{U}_1^{(2i)\top})(I_{p_{d-1}}\otimes[\boldsymbol{\mathcal{Y}}]_{d-2})A^{(p_{d-1},p_d)}\widehat{V}_d^{(2i-1)}\\
=&\Big(I_{p_{d-1}}\otimes\big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots\\
&\qquad(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})[\boldsymbol{\mathcal{Y}}]_{d-2}\big)\Big)\\
&\cdot\Big(I_{p_{d-1}}\otimes(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\Big)A^{(p_{d-1},r_{d-1})}\\
=&\Big(I_{p_{d-1}}\otimes\big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots\\
&\qquad(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\big)\Big)\\
&\cdot A^{(p_{d-1},r_{d-1})}\\
=&\mathrm{Reshape}\Big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\qquad\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}}),r_{d-2}p_{d-1},r_{d-1}\Big).
\end{aligned}
$$

Since the realignment does not change the Frobenius norm, we have

$$
\begin{aligned}
&\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i)}\right\|_{\mathrm{F}}^2\\
\le&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})[\boldsymbol{\mathcal{Y}}]_{d-2}\\
&\cdot(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\Big\|_{\mathrm{F}}^2. \qquad (33)
\end{aligned}
$$

By similar proof of (33), we have

$$\begin{aligned}
&\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i)}\right\|_{\mathrm{F}}^2\\
\leq&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\widehat{V}_{d-1\perp}^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
\leq&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
=&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2\\
&-\Big\|(I_{p_{d-2}}\otimes\widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-2}}\otimes\widehat{U}_1^{(2i)\top})\\
&\cdot[\boldsymbol{\mathcal{Y}}]_{d-2}(\widehat{V}_d^{(2i-1)}\otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
\leq&\cdots\\
\leq&\|[\boldsymbol{\mathcal{Y}}]_1\|_{\mathrm{F}}^2-\Big\|[\boldsymbol{\mathcal{Y}}]_1(\widehat{V}_d^{(2i-1)}\otimes I_{p_2\ldots p_{d-1}})\cdots\\
&\qquad(\widehat{V}_3^{(2i-1)}\otimes I_{p_2})\widehat{V}_2^{(2i-1)}\Big\|_{\mathrm{F}}^2\\
=&\Big\|[\boldsymbol{\mathcal{Y}}]_1\big(I_{p_2\cdots p_d}-\\
&\qquad P_{(\widehat{V}_d^{(2i-1)}\otimes I_{p_2\ldots p_{d-1}})\cdots(\widehat{V}_3^{(2i-1)}\otimes I_{p_2})\widehat{V}_2^{(2i-1)}}\big)\Big\|_{\mathrm{F}}^2\\
=&\left\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{X}}}^{(2i-1)}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Similarly, we can prove (11) holds for $k=2i, i\geq 0$.

C. Proof of Theorem 3

Without loss of generality, we assume $\sigma^2=1$. We still let $\widehat{U}_i$, $\widehat{V}_i$, R_i and $\widetilde{R}_i$ denote $\widehat{U}_i^{(0)}$, $\widehat{V}_i^{(1)}$, $R_i^{(0)}$ and $\widetilde{R}_i^{(0)}$, respectively.

Lemma 5 Part 4 immediately shows that (15) holds with probability at least $1-Ce^{-cp}$. Next, we show that with probability at least $1-Ce^{-cp}$,

$$\begin{aligned}
&\left\|\sin\Theta(\widehat{U}_k,\widetilde{U}_k)\right\|\\
\leq&C\frac{\sqrt{\sum_{i=1}^{k-1}p_ir_{i-1}r_i}+\sqrt{p_kr_{k-1}}+\sqrt{p_{k+1}\cdots p_d}}{\lambda_k}\\
\leq&\frac{1}{2},\quad\forall 1\leq k\leq d-1.
\end{aligned}\tag{34}$$

Recall that

$$\widehat{U}_1=\mathrm{SVD}_{r_1}^L([\boldsymbol{\mathcal{Y}}]_1),\quad[\boldsymbol{\mathcal{Y}}]_1=[\boldsymbol{\mathcal{X}}]_1+[\boldsymbol{\mathcal{Z}}]_1,$$

where $[\boldsymbol{\mathcal{X}}]_1\in\mathbb{R}^{p_1\times p_{-1}}$ satisfying $\mathrm{rank}([\boldsymbol{\mathcal{X}}]_1)=r_1$, $[\boldsymbol{\mathcal{Z}}]_1\in\mathbb{R}^{p_1\times p_{-1}}$, by Lemmas 6 and 5, with probability $1-Ce^{-cp}$, we have

$$\|\widehat{U}_{1\perp}^\top[\boldsymbol{\mathcal{X}}]_1\|\leq 2\|[\boldsymbol{\mathcal{Z}}]_1\|\leq C(p_1^{1/2}+(p_2\cdots p_d)^{1/2}).$$

Therefore, with probability at least $1-Ce^{-cp}$,

$$\begin{aligned}
\left\|\sin\Theta(\widehat{U}_1,U_1)\right\|\leq&\frac{\left\|\widehat{U}_{1\perp}^\top U_1U_1^\top[\boldsymbol{\mathcal{X}}]_1\right\|}{s_{r_1}(U_1^\top[\boldsymbol{\mathcal{X}}]_1)}=\frac{\left\|\widehat{U}_{1\perp}^\top[\boldsymbol{\mathcal{X}}]_1\right\|}{s_{r_1}([\boldsymbol{\mathcal{X}}]_1)}\\
\leq&C\frac{\sqrt{p_1}+\sqrt{p_2\cdots p_d}}{\lambda_1}.
\end{aligned}$$

For $2\leq i\leq j\leq d-1$, by the definition of $\widetilde{U}_i$ and Lemma 2, we have

$$\begin{aligned}
[\boldsymbol{\mathcal{X}}]_j=&(I_{p_2\cdots p_j}\otimes[\boldsymbol{\mathcal{X}}]_1)A^{(p_2\cdots p_j,p_{j+1}\cdots p_d)}\\
=&(I_{p_2\cdots p_j}\otimes(P_{U_1}[\boldsymbol{\mathcal{X}}]_1))A^{(p_2\cdots p_j,p_{j+1}\cdots p_d)}\\
=&(I_{p_2\cdots p_j}\otimes P_{U_1})(I_{p_2\cdots p_j}\otimes[\boldsymbol{\mathcal{X}}]_1)A^{(p_2\cdots p_j,p_{j+1}\cdots p_d)}\\
=&(I_{p_2\cdots p_j}\otimes U_1)(I_{p_2\cdots p_j}\otimes U_1^\top)[\boldsymbol{\mathcal{X}}]_j
\end{aligned}\tag{35}$$

and

$$\begin{aligned}
&\left(I_{p_i\cdots p_j}\otimes\widehat{U}_{i-1}^\top\right)\cdots\left(I_{p_2\cdots p_j}\otimes\widehat{U}_1^\top\right)[\boldsymbol{\mathcal{X}}]_j\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes(I_{p_i}\otimes\widehat{U}_{i-1}^\top)\right)\cdots\\
&\left(I_{p_{i+1}\cdots p_j}\otimes(I_{p_2\cdots p_i}\otimes\widehat{U}_1^\top)\right)\\
[4pt]&\cdot(I_{p_{i+1}\cdots p_j}\otimes[\boldsymbol{\mathcal{X}}]_i)A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\left((I_{p_i}\otimes\widehat{U}_{i-1}^\top)\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_i\right)\right)\\
&\cdot A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\left(P_{\widetilde{U}_i}(I_{p_i}\otimes\widehat{U}_{i-1}^\top)\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_i\right)\right)\\
&\cdot A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes P_{\widetilde{U}_i}\right)\\
&\cdot\left(I_{p_{i+1}\cdots p_j}\otimes\left((I_{p_i}\otimes\widehat{U}_{i-1}^\top)\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_i\right)\right)\\
&\cdot A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\widetilde{U}_i\right)\left(I_{p_{i+1}\cdots p_j}\otimes\widetilde{U}_i^\top\right)\left(I_{p_i\cdots p_j}\otimes\widehat{U}_{i-1}^\top\right)\cdots\\
&\left(I_{p_2\cdots p_j}\otimes\widehat{U}_1^\top\right)[\boldsymbol{\mathcal{X}}]_j,
\end{aligned}\tag{36}$$

where $I_{p_{i+1}\cdots p_j}=1$ if $i=j$. Let

$$L_k=\left\|\sin\Theta\left(\widetilde{U}_k,\widehat{U}_k\right)\right\|,\quad 2\leq k\leq d-1.$$

For $k=2$, by (35) and Lemma 4, with probability at least $1-Ce^{-cp}$,

$$\begin{aligned}
&s_{r_2}\left((I_{p_2}\otimes\widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\right)\\
\geq&s_{\min}\left((I_{p_2}\otimes\widehat{U}_1^\top)(I_{p_2}\otimes U_1)\right)s_{r_2}([\boldsymbol{\mathcal{X}}]_2)
\end{aligned}$$

$$
\begin{aligned}
&= s_{\min}(\widehat{U}_1^\top U_1)\lambda_2 \\
&= \sqrt{1 - \|\sin\Theta(\widehat{U}_1, U_1)\|^2}\lambda_2 \\
&\ge \sqrt{\frac{3}{4}}\lambda_2.
\end{aligned}
$$

Since $\widehat{U}_2 = \mathrm{SVD}_{r_2}^L((I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Y}}]_2)$, and $(I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Y}}]_2 = (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2 + (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_2$, by Lemma 6 and Lemma 4, we know that with probability at least $1 - Ce^{-cpr}$,

$$
\begin{aligned}
&\|\widehat{U}_{2\perp}^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\| \\
\le& 2\|(I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_2\| \\
\le& C(\sqrt{p_2 r_1} + (p_3 \cdots p_d)^{1/2} + \sqrt{p_1 r_1}).
\end{aligned}
$$

Combine the two previous inequalities together and recall that $\widetilde{U}_2$ is the left singular space of $(I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2$, we have

$$
\begin{aligned}
\left\|\sin\Theta\left(\widehat{U}_2, \widetilde{U}_2\right)\right\| &\le \frac{\|\widehat{U}_{2\perp}^\top \widetilde{U}_2 \widetilde{U}_2^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\|}{s_{r_2}\left(\widetilde{U}_2^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\right)} \\
&= \frac{\|\widehat{U}_{2\perp}^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\|}{s_{r_2}\left((I_{p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_2\right)} \\
&\le C\frac{\sqrt{p_1 r_1} + \sqrt{p_2 r_1} + (p_3 \cdots p_d)^{1/2}}{\lambda_2}
\end{aligned}
$$

with probability at least $1 - Ce^{-cp}$.

Assume that (34) holds for $k \le j-1$ with probability $1 - Ce^{-cp}$. For $k = j$, by Lemma 4 and (36), with probability at $1 - Ce^{-cp}$, we have

$$
\begin{aligned}
&s_{r_j}\Big((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top) \\
&\qquad \cdot [\boldsymbol{\mathcal{X}}]_j\Big) \\
\ge& s_{\min}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_j} \otimes \widetilde{U}_{j-1})\right) \\
&\cdot s_{r_j}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right) \\
=& s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) \\
&\cdot s_{r_j}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right) \\
\ge& s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) \\
&\cdot s_{\min}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)(I_{p_{j-1}p_j} \otimes \widetilde{U}_{j-2})\right) \\
&\cdot s_{r_j}\left((I_{p_{j-2}p_{j-1}p_j} \otimes \widehat{U}_{j-3}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right) \\
\ge& \cdots \\
\ge& s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) \cdots s_{\min}\left(\widehat{U}_1^\top \widetilde{U}_1\right) s_{r_j}([\boldsymbol{\mathcal{X}}]_j) \\
=& \sqrt{1 - L_{j-1}^2} \cdots \sqrt{1 - L_1^2}\lambda_j \\
\ge& (\sqrt{3/4})^{j-1}\lambda_j \ge c\lambda_j.
\end{aligned} \tag{37}
$$

In the last inequality, we used the fact that d is a fixed number and $(\sqrt{3/4})^{j-1} \ge (\sqrt{3/4})^{d-1} \ge c$.

By the definition of $\widehat{U}_j$ and Lemma 3, we have

$$
\widehat{U}_j = \mathrm{SVD}_{r_j}^L\Big((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Y}}]_j\Big).
$$

Note that

$$
\begin{aligned}
&(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Y}}]_j \\
=& (I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j \\
&+ (I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_j,
\end{aligned}
$$

by Lemma 6, with probability at least $1 - e^{-cpr^2}$,

$$
\begin{aligned}
&\Big\|\widehat{U}_{j\perp}^\top (I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \\
&\qquad \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\Big\| \\
\le& 2\Big\|(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \\
&\qquad \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{Z}}]_j\Big\| \\
\le& C\left[\left(\sum_{i=1}^{j-1} p_i r_{i-1} r_i\right)^{1/2} + (p_j r_{j-1})^{1/2} + (p_{j+1} \cdots p_d)^{1/2}\right].
\end{aligned}
$$

Therefore, with probability at least $1 - Ce^{-cp}$,

$$
\begin{aligned}
&\left\|\sin\Theta\left(\widehat{U}_j, \widetilde{U}_j\right)\right\| \\
\le& \frac{\left\|\widehat{U}_{j\perp}^\top \widetilde{U}_j \widetilde{U}_j^\top (I_{p_j} \otimes \widehat{U}_{j-1}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right\|}{s_{r_j}\left(\widetilde{U}_j^\top (I_{p_j} \otimes \widehat{U}_{j-1}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right)} \\
=& \frac{\left\|\widehat{U}_{j\perp}^\top (I_{p_j} \otimes \widehat{U}_{j-1}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right\|}{s_{r_j}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_j\right)} \\
\le& C\frac{\left(\sum_{i=1}^{j-1} p_i r_{i-1} r_i\right)^{1/2} + (p_j r_{j-1})^{1/2} + (p_{j+1} \cdots p_d)^{1/2}}{\lambda_j}.
\end{aligned}
$$

Therefore, (13) holds with probability $1 - Ce^{-cp}$.

Finally, we consider (14). Let $\mathcal{E}_0 = \{(13) \text{ and } (15) \text{ hold}\}$. Without loss of generality, we only show that under $\mathcal{E}_0$,

$$
\left\|\sin\Theta\left(\widehat{V}_k, \widetilde{V}_k\right)\right\| \le C\frac{\sqrt{\sum_{i=1}^d p_i r_{i-1} r_i}}{\lambda_{k-1}} \le \frac{1}{2}, \quad \forall 2 \le k \le d. \tag{38}
$$

In fact, (38) can be proved by induction. Let $V_d \in \mathbb{R}^{p_d \times r_{d-1}}$ be the right singular space of $[\boldsymbol{\mathcal{X}}]_{d-1}$. Then there exists an orthogonal matrix $\widetilde{Q}_{d-1} \in \mathbb{O}_{r_{d-1}}$ such that

$$
\begin{aligned}
&V_d \widetilde{Q}_{d-1} \\
=& \mathrm{SVD}^R\big(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \cdots p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{d-1}\big).
\end{aligned}
$$

Similarly to (37), under $\mathcal{E}_0$,

$$
\begin{aligned}
&s_{r_{d-1}}\left(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \cdots p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{d-1}\right) \\
\ge& \left(\sqrt{3/4}\right)^{d-1} \lambda_{d-1} \ge c\lambda_{d-1}.
\end{aligned}
$$

Therefore, by Lemma 6, under $\mathcal{E}_0$,

$$
\begin{aligned}
&\left\|\sin\Theta\left(\widehat{V}_d, V_d\right)\right\| \\
=& \left\|\sin\Theta\left(\widehat{V}_d, V_d \widetilde{Q}_{d-1}\right)\right\| \\
\le& \frac{\left\|\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \cdots p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{d-1} \widehat{V}_{d\perp}^\top\right\|}{s_{r_{d-1}}\left(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \cdots p_2} \otimes \widehat{U}_1^\top)[\boldsymbol{\mathcal{X}}]_{d-1}\right)}
\end{aligned}
$$

$$\leq \frac{2\left\|\widehat{U}_{d-1}^{\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{\top})\cdots(I_{p_{d-1}\cdots p_2}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Z}}]_{d-1}\right\|}{s_{r_{d-1}}\left(\widehat{U}_{d-1}^{\top}(I_{p_{d-1}}\otimes\widehat{U}_{d-2}^{\top})\cdots(I_{p_{d-1}\cdots p_2}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{d-1}\right)}$$
$$\leq C\frac{\sqrt{\sum_{i=1}^d p_i r_{i-1} r_i}}{\lambda_{d-1}}.$$

Suppose (38) holds for $j+1\leq k\leq d$. For $k=j$, since $\widetilde{V}_j$ is the right singular space of $[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})$, there exists $\widetilde{Q}_{j-1}\in\mathbb{O}_{r_{j-1}}$ such that

$$\begin{aligned}&\widetilde{V}_j\widetilde{Q}_{j-1}\\ =&SVD^R\big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})\\ &\cdot[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\big).\end{aligned}$$

By Lemma 4, (35), (36) and (37), under $\mathcal{E}_0$,

$$\begin{aligned}&s_{r_{j-1}}\Big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})\\ &\quad\cdot[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\Big)\\ \geq& s_{r_{j-1}}\Big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\quad\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+2}\otimes I_{p_jp_{j+1}})\\ &\quad\cdot(\widetilde{V}_{j+1}\otimes I_{p_j})\Big)\\ &\cdot s_{\min}\Big((\widetilde{V}_{j+1}^{\top}\otimes I_{p_j})(\widehat{V}_{j+1}\otimes I_{p_j})\Big)\\ =&s_{r_{j-1}}\Big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\quad\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+2}\otimes I_{p_jp_{j+1}})\Big)\\ &\cdot s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\ \geq&\cdots\\ \geq& s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\right)\\ &\cdot s_{\min}(\widetilde{V}_d^{\top}\widehat{V}_d)\cdots s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\ \geq& s_{\min}(\widehat{U}_{j-1}^{\top}\widetilde{U}_{j-1})\\ &\cdot s_{r_{j-1}}\left((I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\right)\\ &\cdot s_{\min}(\widetilde{V}_d^{\top}\widehat{V}_d)\cdots s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\ \geq&\left(\sqrt{\frac{3}{4}}\right)^{j-1}\lambda_{j-1}\cdot\left(\sqrt{\frac{3}{4}}\right)^{d-j}\geq c\lambda_{j-1}.\end{aligned}$$

Note that $\widehat{V}_j\in\mathbb{O}_{p_jr_j,r_{j-1}}$ is the right singular space of $\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Y}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})$ and

$$\begin{aligned}&\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Y}}]_{j-1}\\ &\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\\ =&\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\\ &+\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Z}}]_{j-1}\\ &\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j}),\end{aligned}$$

By Lemma 6, under $\mathcal{E}_0$,

$$\begin{aligned}&\left\|\sin\Theta\left(\widehat{V}_j,\widetilde{V}_j\right)\right\|=\left\|\sin\Theta\left(\widehat{V}_j,\widetilde{V}_j\widetilde{Q}_{j-1}\right)\right\|\\ \leq&\Big\|\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\widehat{V}_{j\perp}\Big\|\\ &\Big/s_{r_{j-1}}\Big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\quad\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\Big)\\ \leq&2\Big\|\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Z}}]_{j-1}\\ &\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\Big\|\\ &\Big/s_{r_{j-1}}\Big(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\\ &\quad\cdot(\widehat{V}_d\otimes I_{p_j\dots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\Big)\\ \leq&C\frac{\left(\sum_{i=1}^d p_ir_ir_{i-1}\right)^{1/2}}{\lambda_{j-1}}.\end{aligned}$$

Therefore, under $\mathcal{E}_0$, (38) holds.

Thus, we have finished the proof of Theorem 3.

D. Proof of Corollary 1

Let $Q=\{(15),(34)\text{ hold}\}$, then $\mathbb{P}(Q^c)\leq C\exp(-cp)$ and

$$\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^2\leq C\sum_{i=1}^d p_ir_ir_{i-1}\quad\text{under } Q.$$

Under Q^c, due to the property of projection matrices, we know that

$$\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}\right\|_{\mathrm{F}}\leq\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\leq\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}+\|\boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}.$$

Moreover,

$$\begin{aligned}&\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^4\leq C\left(\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}\right\|_{\mathrm{F}}^4+\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^4\right)\\ \leq&C\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^4+C\mathbb{E}\|\boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^4\\ \leq&C\exp(4c_0p)+C\mathbb{E}\left(\chi^2_{p_1\cdots p_d}\right)^2\\ \leq&C\exp(4c_0p)+C(p_1\cdots p_d)^2\\ \leq&C\exp(4c_0p)+C\exp(2c_0p)\leq C\exp(4c_0p).\end{aligned}$$

Therefore, we have the following upper bound for the Frobenius norm risk of $\widehat{\boldsymbol{\mathcal{X}}}$:

$$\begin{aligned}&\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2\\ =&\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2 1_Q+\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2 1_{Q^c}\\ \leq&C\sum_{i=1}^d p_ir_ir_{i-1}+\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^4\cdot\mathbb{P}(Q^c)}\\ \leq&C\sum_{i=1}^d p_ir_ir_{i-1}+C\exp\left((4c_0-c)p/2\right).\end{aligned}$$

By selecting $c_0<c/4$, we have

$$\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}}^{(t)}-\boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2\leq C\sum_{i=1}^d p_ir_ir_{i-1}.$$

Therefore, we have finished the proof of Corollary 1.

E. Proof of Theorem 4

Since the i.i.d. Gaussian distribution, $\boldsymbol{\mathcal{Z}} \sim N(0, \sigma^2)$, is a special case of $\mathcal{D}$ and

$$\begin{aligned}&\inf_{\widehat{\boldsymbol{\mathcal{X}}}} \sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda}), D \in \mathcal{D}} \mathbb{E}_{\boldsymbol{\mathcal{Z}} \sim D} \left\| \widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_{\mathrm{F}}^2 \\ \geq &\inf_{\widehat{\boldsymbol{\mathcal{X}}}} \sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda}), \boldsymbol{\mathcal{Z}} \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)} \mathbb{E}_{\boldsymbol{\mathcal{Z}} \sim D} \left\| \widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_{\mathrm{F}}^2,\end{aligned}$$

we only need to focus on the setting that $\boldsymbol{\mathcal{Z}} \sim N(0, \sigma^2)$ while developing the lower bound result.

Without loss of generality, assume $\sigma^2 = 1$. Since d is a fixed number, we only need to show that for any $1 \leq i \leq d$,

$$\inf_{\widehat{\boldsymbol{\mathcal{X}}}} \sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda})} \mathbb{E} \left\| \widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_{\mathrm{F}}^2 \geq c p_i r_i r_{i-1}. \tag{39}$$

Suppose $\boldsymbol{\mathcal{X}}$ can be written as (1), $U_j \in \mathbb{R}^{(p_j r_{j-1}) \times r_j}$ and $V_j \in \mathbb{R}^{(p_j r_j) \times r_{j-1}}$ are reshaped from $\boldsymbol{\mathcal{G}}_j \in \mathbb{R}^{r_{j-1} \times p_j \times r_j}$, $G_1 = U_1$, $G_d = V_d$. For any $1 \leq i \leq d-1$, by Lemma 1, we have

$$\begin{aligned}[\boldsymbol{\mathcal{X}}]_i = &(I_{p_2 \cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1}) U_i V_{i+1}^\top \\ &\cdot \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1} \cdots p_{d-1}}\right).\end{aligned} \tag{40}$$

For all $j \neq i, 1 \leq j \leq d-1$, let $U_j \overset{\text{i.i.d.}}{\sim} N(0,1), V_d \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_{d-1}, V_d$ are all independent. By Lemma 4, for any $1 \leq j \leq d-1$, we have

$$\begin{aligned}&s_{r_j}\left(\left(I_{p_2 \cdots p_j} \otimes U_1\right) \cdots (I_{p_j} \otimes U_{j-1}) U_j\right) \\ \geq &s_{\min}\left(I_{p_2 \cdots p_j} \otimes U_1\right) \cdots s_{\min}(U_j) = s_{r_1}(U_1) \cdots s_{r_j}(U_j).\end{aligned}$$

Similarly,

$$\begin{aligned}&s_{r_j}\left(V_{j+1}^\top \left(V_{j+2}^\top \otimes I_{p_{j+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{j+1} \cdots p_{d-1}}\right)\right) \\ \geq &s_{r_j}(V_{j+1}) \cdots s_{r_{d-1}}(V_d).\end{aligned}$$

Moreover, Lemma 4 Part 1 tells us

$$\begin{aligned}&s_{r_j}\Big(\left(I_{p_2 \cdots p_j} \otimes U_1\right) \cdots (I_{p_j} \otimes U_{j-1}) U_j V_{j+1}^\top \\ &\quad\cdot \left(V_{j+2}^\top \otimes I_{p_{j+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{j+1} \cdots p_{d-1}}\right)\Big) \\ \geq &s_{r_j}\left(\left(I_{p_2 \cdots p_j} \otimes U_1\right) \cdots (I_{p_j} \otimes U_{j-1}) U_j\right) \\ &\quad\cdot s_{r_j}\left(V_{j+1}^\top \left(V_{j+2}^\top \otimes I_{p_{j+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{j+1} \cdots p_{d-1}}\right)\right) \\ \geq &s_{r_1}(U_1) \cdots s_{r_j}(U_j) s_{r_j}(V_{j+1}) \cdots s_{r_{d-1}}(V_d).\end{aligned} \tag{41}$$

Recall that V_j is reshaped from U_j for all $1 \leq j \leq d-1$, by [93][Corollary 5.35], we know that with probability at least $1 - Ce^{-cp}$, for all $1 \leq j \leq d-1, j \neq i$,

$$\begin{aligned}\frac{\sqrt{p_j r_{j-1}}}{4} \leq &\sqrt{p_j r_{j-1}} - \sqrt{r_j} - \frac{\sqrt{p_j r_{j-1}}}{25} \\ \leq &s_{r_j}(U_j) \leq s_1(U_j) \\ \leq &\sqrt{p_j r_{j-1}} + \sqrt{r_j} + \frac{\sqrt{p_j r_{j-1}}}{25} \leq 2\sqrt{p_j r_{j-1}}, \\ \frac{\sqrt{p_j r_j}}{4} \leq &s_{r_{j-1}}(V_j) \leq s_1(V_j) \leq 2\sqrt{p_j r_j}, \\ \text{and} \quad \frac{\sqrt{p_d}}{4} \leq &s_{r_{d-1}}(V_d) \leq s_{r_1}(V_d) \leq 2\sqrt{p_d}.\end{aligned} \tag{42}$$

For a fixed $U_0 \in \mathbb{O}_{p_i r_{i-1}, r_i}$, define the following ball with radius $\varepsilon > 0$,

$$B(U_0, \varepsilon) = \left\{U' \in \mathbb{O}_{p_i r_{i-1}, r_i} : \|\sin \Theta(U', U_0)\|_{\mathrm{F}} \leq \varepsilon\right\}.$$

By Lemma 1 in [94], for $0 < \alpha < 1$ and $0 < \varepsilon \leq 1$, there exist $\widetilde{U}_i^{(1)'}, \ldots, \widetilde{U}_i^{(m)'} \subseteq B(U_0, \varepsilon)$ such that

$$\begin{aligned}&m \geq \left(\frac{c_0}{\alpha}\right)^{r_i(p_i r_{i-1} - r_i)}, \\ &\min_{1 \leq j \neq k \leq m} \left\|\sin \Theta\left(\widetilde{U}_i^{(j)'}, \widetilde{U}_i^{(k)'}\right)\right\|_{\mathrm{F}} \geq \alpha\varepsilon.\end{aligned}$$

By Lemma 1 in [37], one can find a rotation matrix $O_k \in \mathbb{O}_{r_i}$ such that

$$\|U_0 - \widetilde{U}_i^{(k)'} O_k\|_{\mathrm{F}} \leq \sqrt{2} \left\|\sin \Theta\left(U_0, \widetilde{U}_i^{(k)'}\right)\right\|_{\mathrm{F}} \leq \sqrt{2}\varepsilon.$$

Let $\widetilde{U}_i^{(k)} = \widetilde{U}_i^{(k)'} O_k$, we have

$$\begin{aligned}&\left\|\widetilde{U}_i^{(k)} - U_0\right\|_{\mathrm{F}} \leq \sqrt{2}\varepsilon, \\ &\left\|\sin \Theta\left(\widetilde{U}_i^{(j)}, \widetilde{U}_i^{(k)}\right)\right\|_{\mathrm{F}} \geq \alpha\varepsilon, \quad 1 \leq j < k \leq m.\end{aligned}$$

Let $U_i^{(k)} = S + \widetilde{U}_i^{(k)}$, where $S \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$. Set $\tau \geq 8/\sqrt{p_i}$, [93][Corollary 5.35] shows that with probability at least $1 - Ce^{-cp}$,

$$\begin{aligned}\frac{\tau\sqrt{p_i r_{i-1}}}{8} \leq &\tau\left(\sqrt{p_i r_{i-1}} - \sqrt{r_i} - \frac{\sqrt{p_i r_{i-1}}}{25}\right) - 1 \\ \leq &s_{r_i}(S) - s_1\left(\widetilde{U}_i^{(k)}\right) \leq s_{r_i}\left(U_i^{(k)}\right) \\ \leq &s_1\left(U_i^{(k)}\right) \leq s_1(S) + s_1\left(\widetilde{U}_i^{(k)}\right) \\ \leq &\tau\left(\sqrt{p_i r_{i-1}} + \sqrt{r_i} + \frac{\sqrt{p_i r_{i-1}}}{25}\right) + 1 \\ \leq &2\tau\sqrt{p_i r_{i-1}}.\end{aligned} \tag{43}$$

If $2 \leq i \leq d-1$, since $V_i^{(k)}$ is reshaped from $U_i^{(k)}$, we know that $V_i^{(k)} = T + \widetilde{V}_i^{(k)}$, where $T \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$, and $\widetilde{V}_i^{(k)}$ is realigned from $\widetilde{U}_i^{(k)}$. Notice that

$$s_1(\widetilde{V}_i^{(k)}) = \|\widetilde{V}_i^{(k)}\| \leq \|\widetilde{V}_i^{(k)}\|_{\mathrm{F}} = \|\widetilde{U}_i^{(k)}\|_{\mathrm{F}} = r_i,$$

Since $\tau \geq 8/\sqrt{p_i}$, by [93][Corollary 5.35], with probability at least $1 - Ce^{-cp_i r_i}$,

$$\begin{aligned}\frac{\tau\sqrt{p_i r_i}}{8} \leq &\tau\left(\sqrt{p_i r_i} - \sqrt{r_{i-1}} - \frac{\sqrt{p_i r_i}}{25}\right) - \sqrt{r_i} \\ \leq &s_{r_i}(T) - s_1\left(\widetilde{V}_i^{(k)}\right) \leq s_{r_i}\left(V_i^{(k)}\right) \\ \leq &s_1\left(V_i^{(k)}\right) \leq s_1(T) + s_1\left(\widetilde{V}_i^{(k)}\right) \\ \leq &\tau\left(\sqrt{p_i r_i} + \sqrt{r_{i-1}} + \frac{\sqrt{p_i r_i}}{25}\right) + \sqrt{r_i} \\ \leq &2\tau\sqrt{p_i r_i}.\end{aligned} \tag{44}$$

Choose fixed $U_1, \cdots, U_{i-1}, V_{i+1}, \cdots, V_d, S$ such that (42), (43) and (44) hold. Let

$$\begin{aligned}[\boldsymbol{\mathcal{X}}^{(k)}]_i = &(I_{p_2 \cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1}) U_i^{(k)} V_{i+1}^\top \\ &\cdot \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1} \cdots p_{d-1}}\right)\end{aligned} \tag{45}$$

and $\boldsymbol{\mathcal{X}}^{(k)} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is the corresponding tensor. (41), (42), (43) and (44) together show that

$$\sigma_{r_j}([\boldsymbol{\mathcal{X}}^{(k)}]_j) \geq \tau \prod_{k=1}^{j} \frac{\sqrt{p_k r_{k-1}}}{8} \prod_{k=j+1}^{d} \frac{\sqrt{p_k r_k}}{8} = \tau \frac{\sqrt{p_1\cdots p_d r_1 \cdots r_{d-1}}}{C\sqrt{r_j}}. \tag{46}$$

By setting $\tau = \frac{C \max_{1\leq i\leq d-1}\lambda_i \max_{1\leq j\leq d-1}\sqrt{r_j}}{\sqrt{p_1\cdots p_d r_1 \cdots r_{d-1}}} \vee 8\max_{1\leq i\leq d-1}\sqrt{1/p_i}$, we have

$$\sigma_{r_j}\left([X^{(k)}]_j\right) \geq \lambda_j, \quad \forall 1\leq j\leq d-1.$$

For $1 \leq k < j \leq m$,

$$\begin{aligned}
&\|\boldsymbol{\mathcal{X}}^{(k)} - \boldsymbol{\mathcal{X}}^{(j)}\|_{\mathrm{F}}^2 \\
=&\Big\| (I_{p_2\cdots p_i}\otimes U_1)\cdots(I_{p_i}\otimes U_{i-1})\left(U_i^{(k)} - U_i^{(j)}\right) \\
&\cdot V_{i+1}^\top\left(V_{i+2}^\top \otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\Big\|_{\mathrm{F}}^2 \\
\geq& s_{\min}^2\left((I_{p_2\cdots p_i}\otimes U_1)\cdots(I_{p_i}\otimes U_{i-1})\right) \\
&\cdot \Big\|\left(U_i^{(k)} - U_i^{(j)}\right)V_{i+1}^\top\left(V_{i+2}^\top\otimes I_{p_{i+1}}\right)\cdots \\
&\left(V_d^\top\otimes I_{p_{i+1}\cdots p_{d-1}}\right)\Big\|_{\mathrm{F}}^2 \\
=& s_{r_{i-1}}^2\left(\left(I_{p_2\cdots p_{i-1}}\otimes U_1\right)\cdots U_{i-1}\right) \\
&\cdot s_{r_i}^2\left(V_{i+1}^\top\left(V_{i+2}^\top\otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right) \\
&\cdot \left\|U_i^{(k)} - U_i^{(j)}\right\|_{\mathrm{F}}^2 \\
=& s_{r_{i-1}}^2\left(\left(I_{p_2\cdots p_{i-1}}\otimes U_1\right)\cdots U_{i-1}\right) \\
&\cdot s_{r_i}^2\left(V_{i+1}^\top\left(V_{i+2}^\top\otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right) \\
&\cdot \left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)}\right\|_{\mathrm{F}}^2 \\
\geq& s_{r_1}^2(U_1)\cdots s_{r_{i-1}}^2(U_{i-1}) s_{r_i}^2(V_{i+1})\cdots s_{r_{d-1}}^2(V_d) \\
&\cdot \min_{O\in\mathbb{O}_{r_i}}\left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)}O\right\|_{\mathrm{F}}^2 \\
\geq& \prod_{h=1}^{i-1}\frac{p_h r_{h-1}}{16}\prod_{l=i+1}^{d}\frac{p_l r_l}{16}\min_{O\in\mathbb{O}_{r_i}}\left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)}O\right\|_{\mathrm{F}}^2 \\
\geq& \prod_{h=1}^{i-1}\frac{p_h r_{h-1}}{16}\prod_{l=i+1}^{d}\frac{p_l r_l}{16}\left\|\sin\Theta\left(\widetilde{U}_i^{(k)}, \widetilde{U}_i^{(j)}\right)\right\|_{\mathrm{F}}^2 \\
\geq& c\left(\prod_{h=1}^{i-1}p_h r_{h-1}\prod_{l=i+1}^{d}p_l r_l\right)\alpha^2\varepsilon^2.
\end{aligned}$$

In addition, let $\boldsymbol{\mathcal{Y}}^{(k)} = \boldsymbol{\mathcal{X}}^{(k)} + \boldsymbol{\mathcal{Z}}^{(k)}$ and $\boldsymbol{\mathcal{Z}}^{(k)} \overset{\text{i.i.d.}}{\sim} N(0,1)$. The KL-divergence between distributions $\boldsymbol{\mathcal{Y}}^{(k)}$ and $\boldsymbol{\mathcal{Y}}^{(j)}$ is

$$\begin{aligned}
&D_{KL}\left(\boldsymbol{\mathcal{Y}}^{(k)}\|\boldsymbol{\mathcal{Y}}^{(j)}\right) = \frac{1}{2}\|\boldsymbol{\mathcal{X}}^{(k)} - \boldsymbol{\mathcal{X}}^{(j)}\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\Big\|(I_{p_2\cdots p_i}\otimes U_1)\cdots(I_{p_i}\otimes U_{i-1})\left(U_i^{(k)} - U_i^{(j)}\right) \\
&\cdot V_{i+1}^\top\left(V_{i+2}^\top\otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{i+1}\cdots p_{d-1}}\right)\Big\|_{\mathrm{F}}^2 \\
\leq&\frac{1}{2}\|(I_{p_2\cdots p_i}\otimes U_1)\cdots(I_{p_i}\otimes U_{i-1})\|^2 \\
&\cdot\left\|V_{i+1}^\top\left(V_{i+2}^\top\otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right\|^2 \\
&\cdot\left\|U_i^{(k)} - U_i^{(j)}\right\|_{\mathrm{F}}^2 \\
\leq&\frac{1}{2}s_1^2(U_1)\cdots s_1^2(U_{i-1})s_1^2(V_{i+1})\cdots s_1^2(V_d)\left\|U_i^{(k)} - U_i^{(j)}\right\|_{\mathrm{F}}^2 \\
\leq&\frac{1}{2}\prod_{h=1}^{i-1}(4p_h r_{h-1})\prod_{l=i+1}^{d}(4p_l r_l) \\
&\cdot\left(\left\|U_i^{(k)} - U_0\right\|_{\mathrm{F}} + \left\|U_i^{(k)} - U_0\right\|_{\mathrm{F}}\right)^2 \\
\leq& C\left(\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)\right)\varepsilon^2.
\end{aligned}$$

By generalized Fano's Lemma,

$$\begin{aligned}
&\inf_{\widehat{\boldsymbol{\mathcal{X}}}}\sup_{\boldsymbol{\mathcal{X}}\in\{\boldsymbol{\mathcal{X}}^{(k)}\}_{k=1}^m}\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}} \\
\geq& c\sqrt{\prod_{h=1}^{i-1}p_h r_{h-1}\prod_{l=i+1}^{d}p_l r_l}\,\alpha\varepsilon \\
&\cdot\left(1 - \frac{C\left(\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)\right)\varepsilon^2 + \log 2}{r_i(p_i r_{i-1} - r_i)\log(c_0/\alpha)}\right).
\end{aligned}$$

By setting $\varepsilon = c'\sqrt{\frac{r_i(p_i r_{i-1} - r_i)}{C\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)}} \leq \frac{1}{2}, \alpha = (c_0\wedge 1)/8$, we know that for any $1 \leq i \leq d-1$,

$$\begin{aligned}
&\inf_{\widehat{\boldsymbol{\mathcal{X}}}}\sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda})}\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2 \\
\geq&\left(\inf_{\widehat{\boldsymbol{\mathcal{X}}}}\sup_{\boldsymbol{\mathcal{X}}\in\{\boldsymbol{\mathcal{X}}^{(k)}\}_{k=1}^m}\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}\right)^2 \\
\geq& c_1 r_i p_i r_{i-1}.
\end{aligned}$$

For $i = d$, similarly to the case $i = 1$, we have

$$\inf_{\widehat{\boldsymbol{\mathcal{X}}}}\sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\boldsymbol{\lambda})}\mathbb{E}\left\|\widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_{\mathrm{F}}^2 \geq c_1 p_d r_{d-1}.$$

Therefore, we have proved Theorem 4.

F. Proof of Proposition 1

Define $\widetilde{G}_1 \in \mathbb{R}^{p\times r_1}, \widetilde{\boldsymbol{\mathcal{G}}}_k \in \mathbb{R}^{r_{k-1}\times p\times r_k}$, $\widetilde{G}_d \in \mathbb{R}^{p\times r_{d-1}}$ such that

$$\widetilde{G}_{1,[i,l]} = (G_1(i))_l, \quad \forall i\in[p], l\in[r_1],$$

$$\widetilde{\boldsymbol{\mathcal{G}}}_{k,[j,i,l]} = \left(G_k(i, e_j^{(r_{k-1})})\right)_l, \quad \forall i\in[p], j\in[r_{k-1}], l\in[r_k], \quad 2\leq k\leq d-1,$$

$$\widetilde{G}_{d,[i,l]} = G_d(i, e_l^{(r_{d-1})}), \forall i\in[p], l\in[r_{d-1}]$$

where $e_i^{(k)}$ is the i-th canonical basis of $\mathbb{R}^k$. Then

$$\widetilde{P}_1(X_{t+1}) = \widetilde{G}_{1,[X_{t+1},:]}^\top \in \mathbb{R}^{r_1},$$

$$\widetilde{P}_2(X_{t+1}, X_{t+2}) = G_2\left(X_{t+2}, \widetilde{P}_1(X_{t+1})\right) \overset{\text{linear map}}{=} \sum_{j=1}^{r_1} G_2(X_{t+2}, e_j^{(r_1)})\left(\widetilde{P}_1(X_{t+1})\right)_j = \left(\widetilde{G}_{1,[X_{t+1},:]}\widetilde{\mathcal{G}}_{2,[:,X_{t+2},:]}\right)^\top.$$

By induction, for any $2 \le k \le d-1$,

$$\begin{aligned}&\widetilde{P}_k(X_{t+1},\dots,X_{t+k})\\ &= G_k(X_{t+k}, \widetilde{P}_{k-1}(X_{t+1},\dots,X_{t+k-1}))\\ &\overset{\text{linear map}}{=} \sum_{j=1}^{r_{k-1}} G_k(X_{t+k}, e_j^{(r_{k-1})})\left(\widetilde{P}_{k-1}(X_{t+1},\dots,X_{t+k-1})\right)_j\\ &= \widetilde{\mathcal{G}}_{k,[:,X_{t+k},:]}^\top \widetilde{P}_{k-1}(X_{t+1},\dots,X_{t+k-1})\\ &= \left(\widetilde{G}_{1,[X_{t+1},:]}\widetilde{\mathcal{G}}_{2,[:,X_{t+2},:]}\cdots\widetilde{\mathcal{G}}_{k,[:,X_{t+k},:]}\right)^\top\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left(X_{t+d}|X_{t+1},\dots,X_{t+d-1}\right)\\ &=G_d(X_{t+d}, \widetilde{P}_{d-1}(X_{t+1},\dots,X_{t+d-1}))\\ &=\widetilde{P}_{d-1}^\top(X_{t+1},\dots,X_{t+d-1})\widetilde{G}_{d,[X_{t+d},:]}^\top\\ &=\widetilde{G}_{1,[X_{t+1},:]}\widetilde{\mathcal{G}}_{2,[:,X_{t+2},:]}\cdots\widetilde{\mathcal{G}}_{d-1,[:,X_{t+d-1},:]}\widetilde{G}_{d,[X_{t+d},:]}^\top.\end{aligned}$$

Therefore,

$$\boldsymbol{\mathcal{P}} = [\![\widetilde{G}_1, \widetilde{\mathcal{G}}_2, \dots, \widetilde{\mathcal{G}}_{d-1}, \widetilde{G}_d]\!]$$

and has TT-rank $(r_1, \dots, r_{d-1})$.

G. Proof of Proposition 2

Let $\boldsymbol{\mathcal{Z}} = \widehat{\boldsymbol{\mathcal{P}}}^{\text{emp}} - \boldsymbol{\mathcal{P}}$, then $\mathbb{E}\boldsymbol{\mathcal{Z}} = 0$. Let

$$\boldsymbol{\mathcal{T}}_{i_1,\dots,i_d}^{(k)} = 1_{\{X(i_1,\dots,i_{d-1};k)=i_d\}}, \quad \forall 1 \le k \le n; \quad 1 \le i_1,\dots,i_d \le p$$

and

$$\boldsymbol{\mathcal{Z}}_{i_1,\dots,i_d}^{(k)} = \boldsymbol{\mathcal{T}}_{i_1,\dots,i_d}^{(k)} - \mathbb{P}\left(i_d|i_1,\dots,i_{d-1}\right), \quad \forall 1 \le k \le n; 1 \le i_1,\dots,i_d \le p.$$

Then $\mathbb{E}\boldsymbol{\mathcal{Z}}^{(k)} = 0$. Moreover, by definition, for any $1 \le j \le d-1$, the rows of $\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j \in \mathbb{R}^{p^j \times p^{d-j}}$ are independent, and there exists a partition $\{\Omega_1^{(j)}, \dots, \Omega_{p^{d-j-1}}^{(j)}\}$ of $\{1,\dots,p^{d-j}\}$ satisfying $\left|\Omega_1^{(j)}\right| = \cdots = \left|\Omega_{p^{d-j-1}}^{(j)}\right| = p$, such that $\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{[:,\Omega_1^{(j)}]}, \dots, \left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{[:,\Omega_{p^{d-j-1}}^{(j)}]}$ are independent and

$$\sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{T}}^{(k)}\right]_j\right)_{m,l} = 1, \quad \forall 1 \le m \le p^j, 1 \le k \le n.$$

Therefore,

$$\begin{aligned}&\sum_{l\in\Omega_i^{(j)}}\left|\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{m,l}\right|\\ &\le \sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{T}}^{(k)}\right]_j\right)_{m,l} + \mathbb{E}\sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{T}}^{(k)}\right]_j\right)_{m,l}\\ &=2, \quad \forall 1 \le m \le p^j, 1 \le k \le n.\end{aligned}$$

For any fixed $x_1 \in \mathbb{R}^{p^j}$ and $x_2 \in \mathbb{R}^{p^{d-j}}$ satisfying $\|x_1\|_2 = 1$ and $\|x\|_2 = 1$, we have

$$\begin{aligned}&\left|\sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{m,l}(x_2)_l\right|\\ &\le \max_{l\in\Omega_i^{(j)}}(x_2)_l \sum_{l\in\Omega_i^{(j)}}\left|\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{m,l}\right|\\ &\le 2\max_{l\in\Omega_i^{(j)}}(x_2)_l \le 2\left\|(x_2)_{\Omega_i^{(j)}}\right\|_2.\end{aligned}$$

By [95, Exercise 2.4], $\sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{m,l}(x_2)_l$ is $2\left\|(x_2)_{\Omega_i^{(j)}}\right\|_2$-sub-Gaussian. Therefore,

$$x_1^\top\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j x_2 = \sum_{m=1}^{p^j}(x_1)_m \sum_{i=1}^{p^{d-j-1}}\left(\sum_{l\in\Omega_i^{(j)}}\left(\left[\boldsymbol{\mathcal{Z}}^{(k)}\right]_j\right)_{m,l}(x_2)_l\right)$$

is $\left(\sum_{m=1}^{p^j}(x_1)_m^2\sum_{i=1}^{p^{d-j-1}} 4\left\|(x_2)_{\Omega_i^{(j)}}\right\|_2^2\right)^{1/2} = 2\|x_1\|_2\|x_2\|_2 = 2$-sub-Gaussian. Notice that $\boldsymbol{\mathcal{Z}} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{\mathcal{Z}}^{(k)}$, the Hoeffding bound [95, Proposition 2.5] shows that

$$\mathbb{P}\left(\left|x_1^\top[\boldsymbol{\mathcal{Z}}]_j x_2\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{8}\right), \quad \forall t \ge 0.$$

Therefore, for any fixed $U \in \mathbb{O}_{p^j,r_j}, V \in \mathbb{O}_{p^{d-j},pr_{j+1}}$, $x \in \mathbb{R}^{r_j}, y \in \mathbb{R}^{pr_{j+1}}$ with $\|x\|_2 = 1$ and $\|y\|_2 = 1$,

$$\mathbb{P}\left(\left|x^\top U^\top[\boldsymbol{\mathcal{Z}}]_j V^\top y\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{8}\right), \quad \forall t \ge 0.$$

Similarly to the proof of (49), with probability at least $1 - Ce^{-cp}$, for all $1 \le k \le d-1$,

$$\begin{aligned}&\left\|\widehat{U}_k^{(0)\top}(I_p \otimes \widehat{U}_{k-1}^{(0)\top})\cdots(I_{p^{k-1}} \otimes \widehat{U}_1^{(0)\top})[\boldsymbol{\mathcal{Z}}]_k\right.\\ &\quad\left.\cdot(\widehat{V}_d^{(1)} \otimes I_{p^{d-k-1}})\cdots(\widehat{V}_{k+2}^{(1)} \otimes I_p)\right\|\\ &\le C\sqrt{\frac{\sum_{i=1}^d p_i r_i r_{i-1}}{n}}.\end{aligned}$$

Similarly, with probability at least $1 - Ce^{-cp}$,

$$\left\|[\boldsymbol{\mathcal{Z}}]_1(\widehat{V}_d^{(1)} \otimes I_{p^{d-2}})\cdots(\widehat{V}_3^{(1)} \otimes I_p)\widehat{V}_2^{(1)}\right\|$$

$$\leq C\sqrt{\frac{\sum_{i=1}^{d} p_i r_i r_{i-1}}{n}}.$$

Notice that $\|X\|_{\mathrm{F}} \leq \sqrt{r}\|X\|$ if $\mathrm{rank}(X)=r$, by the previous two inequalities and Theorem 1, we know that with probability at least $1-Ce^{-cp}$,

$$\left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)}-\boldsymbol{\mathcal{P}}\right\|_{\mathrm{F}}^2 \leq C\left(\max_{1\leq i\leq d-1} r_i\right)\frac{\sum_{i=1}^{d} p_i r_i r_{i-1}}{n}.$$

Finally, by the definition of $\widehat{\boldsymbol{\mathcal{P}}}$, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{\mathcal{P}}}-\boldsymbol{\mathcal{P}}\right\|_{\mathrm{F}} \leq & \left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)}-\boldsymbol{\mathcal{P}}\right\|_{\mathrm{F}}+\left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)}-\widehat{\boldsymbol{\mathcal{P}}}\right\|_{\mathrm{F}}\\ \leq & 2\left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)}-\boldsymbol{\mathcal{P}}\right\|_{\mathrm{F}},\end{aligned}$$

which has finished the proof of Theorem 2.

H. Proof of Lemma 3

By symmetry, we only need to prove (6). By definition, (6) holds for $k=1$. Suppose it holds for $k=j$. For $k=j+1$, since $S_{j+1} \in \mathbb{R}^{(r_j p_{j+1})\times(p_{j+2}\cdots p_d)}$ is realigned from $\widetilde{S}_j = M_j^\top S_j \in \mathbb{R}^{r_j\times(p_{j+1}\cdots p_d)}$, Lemma 2 that $S_{j+1} = (I_{p_{j+1}} \otimes \widetilde{S}_j)A^{(p_{j+1},p_{j+2}\cdots p_d)}$, where the realignment matrix $A^{(i,j)}$ is defined in (5). Therefore,

$$\begin{aligned}S_{j+1} =& \left(I_{p_{j+1}} \otimes \widetilde{S}_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\ =& \left(I_{p_{j+1}} \otimes M_j^\top S_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\ =& \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_{j+1}} \otimes S_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\ =& \left(I_{p_{j+1}} \otimes M_j^\top\right)\\ &\cdot\left(I_{p_{j+1}} \otimes \left((I_{p_j}\otimes M_{j-1}^\top)\cdots(I_{p_2\cdots p_j}\otimes M_1^\top)[\boldsymbol{\mathcal{T}}]_j\right)\right)\\ &\cdot A^{(p_{j+1},p_{j+2}\cdots p_d)}\\ =& \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_{j+1}} \otimes (I_{p_j}\otimes M_{j-1}^\top)\right)\cdots\\ &\left(I_{p_{j+1}} \otimes (I_{p_2\cdots p_j}\otimes M_1^\top)\right)\left(I_{p_{j+1}} \otimes [\boldsymbol{\mathcal{T}}]_j\right)\\ &\cdot A^{(p_{j+1},p_{j+2}\cdots p_d)}\\ =& \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_j p_{j+1}} \otimes M_{j-1}^\top\right)\cdots\\ &\left(I_{p_2\cdots p_{j+1}} \otimes M_1^\top\right)[\boldsymbol{\mathcal{T}}]_{j+1}.\end{aligned}$$

The third equation and the fifth equation hold since $(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$; the last equation holds since $Y_{j+1} = \left(I_{p_{j+1}}\otimes Y_j\right)A^{(p_{j+1},p_{j+2}\cdots p_d)}$ and $A\otimes(B\otimes C) = (A\otimes B)\otimes C$.

Also notice that $\widetilde{S}_k = M_k^\top S_k$, we have finished the proof of (6).

I. Technical Lemmas

We collect the additional technical lemmas in this section.

Lemma 4:

(1) Suppose $A\in\mathbb{R}^{m_1\times m_2}, B\in\mathbb{R}^{m_2\times m_3}$, where $m_1\geq m_2$. Then

$$s_{\min\{m_2,m_3\}}(AB) \geq s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).$$

(2) Suppose $A\in\mathbb{R}^{m\times p_1}, B\in\mathbb{R}^{n\times p_2}, X\in\mathbb{R}^{p_1\times p_2}$, $\mathrm{rank}(X)=r, p_1\geq m, p_2\geq n$. If $X=U_1MV_1^\top$, where $U_1\in\mathbb{O}_{p_1,m}$ and $V_1\in\mathbb{O}_{p_2,n}$, then

$$\sigma_r(AXB)\geq s_{\min}(AU_1)\sigma_r(X)s_{\min}(V_1^\top B).$$

Proof of Lemma 4: (1) Consider the SVD decomposition $A=U_A\Sigma_A V_A^\top, B=U_B\Sigma_B V_B^\top$, where $U_A\in\mathbb{O}_{m_1,m_2}, V_A\in\mathbb{O}_{m_2}, U_B\in\mathbb{O}_{m_2,\min\{m_2,m_3\}}, V_B\in\mathbb{O}_{\min\{m_2,m_3\},m_3}$, $\Sigma_A=\mathrm{diag}(\sigma_1(A),\ldots,s_{m_2}(A))$ and $\Sigma_B=\mathrm{diag}(s_1(B),\ldots,s_{\min\{m_2,m_3\}}(B))$ are diagonal matrices with nonnegative diagonal entries. Then

$$\begin{aligned}s_{\min\{m_2,m_3\}}(AB) =& s_{\min\{m_2,m_3\}}(U_A\Sigma_A V_A^\top U_B\Sigma_B V_B^\top)\\ =& s_{\min\{m_2,m_3\}}(\Sigma_A V_A^\top U_B\Sigma_B).\end{aligned}$$

For any $x\in\mathbb{R}^{\min\{m_2,m_3\}}$ satisfying $\|x\|_2=1$, we have

$$\begin{aligned}\|\Sigma_A V_A^\top U_B\Sigma_B x\|_2 \geq& s_{m_2}(A)\|V_A^\top U_B\Sigma_B x\|_2\\ =& s_{m_2}(A)\|\Sigma_B x\|_2\\ \geq& s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).\end{aligned}$$

Therefore

$$\begin{aligned}s_{\min\{m_2,m_3\}}(AB) =& s_{\min\{m_2,m_3\}}(\Sigma_A V_A^\top U_B\Sigma_B)\\ \geq& s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).\end{aligned}$$

(2) Consider the SVD decomposition $X=U\Sigma V^\top$, where $U\in\mathbb{O}_{p_1,r}, V\in\mathbb{O}_{p_2,r}$ and Σ is a diagonal matrix. Then we know that there exist two matrices $L\in\mathbb{R}^{m\times r}$ and $R\in\mathbb{R}^{n\times r}$ satisfying $U=U_1L$ and $V=V_1R$. Moreover,

$$\begin{aligned}L^\top L = L^\top U_1^\top U_1 L = U^\top U = I_r,\\ R^\top R = R^\top V_1^\top V_1 R = V^\top V = I_r.\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma_r(AXB) =& \sigma_r(AU_1L\Sigma R^\top V_1^\top B)\\ \geq& s_{\min}(AU_1)\sigma_r(L\Sigma R^\top)s_{\min}(V_1^\top B)\\ =& s_{\min}(AU_1)\sigma_r(X)s_{\min}(V_1^\top B).\end{aligned}$$

□

Lemma 5: Suppose Z is a matrix with independent zero-mean σ-sub-Gaussian entries, d is a fixed number, $r_0=r_d=1$.

(1) Suppose $Z\in\mathbb{R}^{p\times q}$, $A\in\mathbb{R}^{m\times p}, B\in\mathbb{R}^{q\times n}$ satisfy $\|A\|,\|B\|\leq 1$, $m\leq p, n\leq q$. Then

$$\mathbb{P}\left(\|AZB\|\geq 2\sigma\sqrt{m+t}\right)\leq 2\cdot 5^n\exp\left[-c\min\left(\frac{t^2}{m},t\right)\right]. \tag{47}$$

$$\mathbb{P}\left(\|AZB\|_{\mathrm{F}}\geq \sigma\sqrt{mn+t}\right)\leq 2\exp\left[-c\min\left(\frac{t^2}{mn},t\right)\right]. \tag{48}$$

(2) Suppose $Z\in\mathbb{R}^{(p_1\cdots p_k)\times m}$, $2\leq k\leq d-1$. Then

$$\begin{aligned}&\max_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i}\\ \|U_i\|\leq 1}}\left\|(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z\right\|\\ &\leq C\sigma\sqrt{\sum_{i=1}^{k-1}p_i r_{i-1}r_i + p_k r_{k-1} + m}.\end{aligned} \tag{49}$$

with probability at least $1 - C\exp(-c(\sum_{i=1}^{k-1} p_i r_{i-1} r_i + p_k r_{k-1} + m))$.

(3) Suppose $Z \in \mathbb{R}^{(p_1 \cdots p_k)\times(p_{k+1}\cdots p_d)}$, $2 \le k \le d-2$. Then

$$\max_{(U_1,\dots,V_d)\in\mathcal{A}} \big\|U_k^\top (I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z \cdot (V_d \otimes I_{p_{k+1}\dots p_{d-1}})\cdots(V_{k+2}\otimes I_{p_{k+1}})\big\| \le C\sigma\sqrt{\sum_{i=1}^d p_i r_{i-1} r_i} \quad (50)$$

with probability at least $1 - C\exp(-c\sum_{i=1}^d p_i r_{i-1} r_i)$. Here,

$$\mathcal{A} = \{(U_1,\dots,U_k,V_{k+2},\dots,V_d) : U_i \in \mathbb{R}^{(p_i r_{i-1})\times r_i}, \|U_i\| \le 1, V_j \in \mathbb{R}^{(p_i r_i)\times r_{i-1}}, \|V_j\| \le 1\}. \quad (51)$$

(4) Suppose $Z \in \mathbb{R}^{(p_1\cdots p_{d-1})\times p_d}$. Then with probability at least $1 - C\exp(-c\sum_{i=1}^d p_i r_{i-1} r_i)$,

$$\max_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i},\\ \|U_i\|\le 1}} \Big\|U_{d-1}^\top (I_{p_{d-1}}\otimes U_{d-2}^\top)\cdots (I_{p_2\cdots p_{d-1}}\otimes U_1^\top)Z\Big\|_{\mathrm{F}} \le C\sigma\sqrt{\sum_{i=1}^d p_i r_{i-1} r_i}. \quad (52)$$

(5) Suppose $Z \in \mathbb{R}^{(p_1 \cdots p_k)\times(p_{k+1}\cdots p_d)}$, $2 \le k \le d-2$. Then

$$\max_{(U_1,\dots,V_d)\in\mathcal{A}} \big\|U_k^\top (I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z \cdot (V_d \otimes I_{p_{k+1}\dots p_{d-1}})\cdots(V_{k+2}\otimes I_{p_{k+1}})\big\|_{\mathrm{F}} \le C\sigma\sqrt{\sum_{i=1}^d p_i r_{i-1} r_i} \quad (53)$$

with probability at least $1 - C\exp(-c\sum_{i=1}^d p_i r_{i-1} r_i)$. Here, $\mathcal{A}$ is defined in (51).

Proof of Lemma 5: W.O.L.G., assume $\sigma = 1$.
(1) For fixed $x \in \mathbb{R}^n$ satisfying $\|x\|_2 = 1$, we have $AZBx = (x^\top B^\top \otimes A)\mathrm{vec}(Z)$. Since Z_{ij} is 1-sub-Gaussian, we know that $\mathrm{Var}(Z_{ij}) \le 1$. In addition,

$$\begin{aligned}
&\mathbb{E}\|(x^\top B^\top \otimes A)\mathrm{vec}(Z)\|_2^2\\
=&\mathbb{E}\left[\mathrm{tr}\left(\mathrm{vec}(Z)^\top (x^\top B^\top\otimes A)^\top (x^\top B^\top\otimes A)\mathrm{vec}(Z)\right)\right]\\
=&\mathrm{tr}\left[\mathbb{E}\left((x^\top B^\top\otimes A)^\top (x^\top B^\top\otimes A)\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right]\\
=&\mathrm{tr}\left[(x^\top B^\top\otimes A)^\top (x^\top B^\top\otimes A)\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right]\\
\le&\mathrm{tr}\left((x^\top B^\top\otimes A)^\top (x^\top B^\top\otimes A)\right)\\
=&\left\|x^\top B^\top\otimes A\right\|_{\mathrm{F}}^2 = \|Bx\|_2^2\|A\|_{\mathrm{F}}^2 \le \|x\|_2^2\|A\|_{\mathrm{F}}^2\\
\le& m.
\end{aligned} \quad (54)$$

The first inequality holds since $\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)$ is a diagonal matrix with diagonal entries $\mathrm{Var}(Z_{ij}) \le 1$; the last inequality is due to $\|A\|_{\mathrm{F}} \le \min\{m,p\}\|A\|_2 \le m$.

By Hanson-Wright inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\|AZBx\|_2^2 - m \ge t\right)\\
\le& 2\exp\Big[-c\min\Big(\frac{t^2}{\|(Bxx^\top B^\top)\otimes(A^\top A)\|_{\mathrm{F}}^2}, \frac{t}{\|(Bxx^\top B^\top)\otimes(A^\top A)\|}\Big)\Big].
\end{aligned}$$

Since $\|x\|_2 = 1$ and $\|A\|, \|B\| \le 1$,

$$\begin{aligned}
&\|(Bxx^\top B^\top)\otimes(A^\top A)\|_{\mathrm{F}}^2\\
=&\|Bxx^\top B^\top\|_{\mathrm{F}}^2\|A^\top A\|_{\mathrm{F}}^2 = (x^\top B^\top Bx)^2\|A^\top A\|_{\mathrm{F}}^2\\
\le&(x^\top x)^2\|A^\top A\|_{\mathrm{F}}^2 = \sum_{i=1}^{\min\{m,p\}}\sigma_i^4(A) \le m,\\
&\|(Bxx^\top B^\top)\otimes(A^\top A)\| \le \|Bxx^\top B^\top\|\|A^\top A\|\\
\le&\|xx^\top\|\|A^\top A\| \le 1.
\end{aligned}$$

Thus, for fixed x satisfying $\|x\|_2 = 1$, we have

$$\mathbb{P}\left(\|AZBx\|_2^2 \ge m + t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right]. \quad (55)$$

By [93][Lemma 5.2], there exists $\mathcal{N}_{1/2}$, a 1/2-net of $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, such that $|\mathcal{N}_{1/2}| \le 5^n$. The union bound, [93][Lemma 5.2] and (55) together imply that

$$\begin{aligned}
&\mathbb{P}\left(\|AZB\| \ge 2\sqrt{m+t}\right)\\
\le&\mathbb{P}\left(\max_{x\in\mathcal{N}_{1/2}}\|AZBx\|_2 \ge \sqrt{m+t}\right)\\
\le&2\cdot 5^n\exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right].
\end{aligned}$$

For $\|AZB\|_{\mathrm{F}}$, note that $AZB = (B^\top\otimes A)\mathrm{vec}(Z)$, Similarly to (54), we have

$$\begin{aligned}
&\mathbb{E}\|(B^\top\otimes A)\mathrm{vec}(Z)\|_2^2\\
=&\mathbb{E}\left[\mathrm{vec}(Z)^\top (B^\top\otimes A)^\top (B^\top\otimes A)\mathrm{vec}(Z)\right]\\
=&\mathbb{E}\left\{\mathrm{tr}\left[\mathrm{vec}(Z)^\top (B^\top\otimes A)^\top (B^\top\otimes A)\mathrm{vec}(Z)\right]\right\}\\
=&\mathrm{tr}\left\{\mathbb{E}\left[(B^\top\otimes A)^\top (B^\top\otimes A)\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right]\right\}\\
=&\mathrm{tr}\left[(B^\top\otimes A)^\top (B^\top\otimes A)\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right]\\
\le&\mathrm{tr}\left[(B^\top\otimes A)^\top (B^\top\otimes A)\right]\\
=&\|B^\top\otimes A\|_{\mathrm{F}}^2 = \|B\|_{\mathrm{F}}^2\|A\|_{\mathrm{F}}^2\\
\le& mn.
\end{aligned}$$

By Hanson-Wright inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\|AZB\|_{\mathrm{F}}^2 - mn \ge t\right)\\
\le&2\exp\Big[-c\min\Big(\frac{t^2}{\|(BB^\top)\otimes(A^\top A)\|_{\mathrm{F}}^2}, \frac{t}{\|(BB^\top)\otimes(A^\top A)\|}\Big)\Big].
\end{aligned}$$

Since $\|A\|, \|B\| \le 1$, we have

$$\begin{aligned}
&\|(BB^\top)\otimes(A^\top A)\|_{\mathrm{F}} = \sqrt{\|A^\top A\|_{\mathrm{F}}^2\|BB^\top\|_{\mathrm{F}}^2}\\
=&\sqrt{\sum_{i=1}^{\min\{m,p\}}\sigma_i^4(A)\sum_{i=1}^{\min\{q,n\}}\sigma_i^4(B)} \le \sqrt{mn},\\
&\|(BB^\top)\otimes(A^\top A)\| \le 1.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\|AZB\|_{\mathrm{F}}^2 \geq mn + t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{mn}, t\right)\right].$$

(2) For fixed $x \in \mathbb{R}^m$ and $A \in \mathbb{R}^{(p_k r_{k-1})\times(p_1\cdots p_k)}$ satisfying $\|x\|_2 = 1$ and $\|A\| \leq 1$, by (47) with $B = I_m$, we have

$$\begin{aligned}&\mathbb{P}\left(\|AZ\| \geq 2\sqrt{p_k r_{k-1} + t}\right)\\ \leq& 2\cdot 5^m \exp\left[-c\min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right]. \end{aligned} \tag{56}$$

By [48][Lemma 7], for $1 \leq i \leq k-1$, there exist ε-nets: $U_i^{(1)}, \ldots, U_i^{(N_i)} \in \mathbb{R}^{(p_i r_{i-1})\times r_i}$ (here $r_0 = 1$), $N_i \leq ((2+\varepsilon)/\varepsilon)^{(p_i r_{i-1})\times r_i}$, such that

$$\begin{aligned}&\forall U \in \mathbb{R}^{(p_i r_{i-1})\times r_i} \text{ satisfying } \|U\| \leq 1,\\ &\exists 1 \leq j \leq N_i \text{ s.t. } \|U_i^{(j)} - U\| \leq \varepsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\Big(\max_{i_1,\ldots,i_{k-1}} \left\|(I_{p_k}\otimes U_{k-1}^{(i_{k-1})\top})\cdots(I_{p_2\cdots p_k}\otimes U_1^{(i_1)\top})Z\right\|\\ &\geq 2\sqrt{p_k r_{k-1} + t}\Big)\\ \leq& 2((2+\varepsilon)/\varepsilon)^{\sum_{i=1}^{k-1} p_i r_{i-1} r_i} 5^m \exp\left[-c\min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right].\end{aligned} \tag{57}$$

Let

$$\begin{aligned}&U_1^*, \ldots, U_{k-1}^*\\ \in& \mathop{\arg\max}_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i},\\ \|U_i\|\leq 1,\quad 1\leq i\leq k-1}} \left\|(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z\right\|,\\ &M\\ =& \max_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i},\\ \|U_i\|\leq 1,\quad 1\leq i\leq k-1}} \left\|(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z\right\|.\end{aligned}$$

Then for any $1 \leq i \leq k-1$, there exists $1 \leq j_i \leq N_i$, such that $\|U_i^{(j_i)} - U_i^*\| \leq \varepsilon$. Then

$$\begin{aligned}&M\\ =&\left\|(I_{p_k}\otimes U_{k-1}^{*\top})\cdots(I_{p_2\cdots p_k}\otimes U_1^{*\top})Z\right\|\\ \leq&\left\|(I_{p_k}\otimes U_{k-1}^{(j_{k-1})\top})\cdots(I_{p_2\cdots p_k}\otimes U_1^{(j_1)\top})Z\right\|\\ &+\left\|\left(I_{p_k}\otimes(U_{k-1}^* - U_{k-1}^{(j_{k-1})})\right)^\top(I_{p_{k-1}p_k}\otimes U_{k-2}^{(j_{k-2})\top})\cdots\right.\\ &\left.(I_{p_2\cdots p_k}\otimes U_1^{(j_1)\top})Z\right\|\\ &+\cdots\\ &+\left\|(I_{p_k}\otimes U_{k-1}^{*\top})\cdots(I_{p_3\cdots p_k}\otimes U_2^{*\top})\right.\\ &\left.\cdot\left(I_{p_2\cdots p_k}\otimes(U_1^* - U_1^{(j_1)})^\top\right)Z\right\|\\ \leq&\left\|(I_{p_k}\otimes U_{k-1}^{(j_{k-1})\top})\cdots(I_{p_2\cdots p_k}\otimes U_1^{(j_1)\top})Z\right\| + \varepsilon(k-1)M.\end{aligned} \tag{58}$$

Combine (57) and the previous inequality together, we have

$$\begin{aligned}&\mathbb{P}\left(M \geq \frac{2\sqrt{p_k r_{k-1} + t}}{1-(k-1)\varepsilon}\right)\\ \leq& 2((2+\varepsilon)/\varepsilon)^{\sum_{i=1}^{k-1} p_i r_{i-1} r_i} 5^m \exp\left[-c\min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right].\end{aligned} \tag{59}$$

By setting $\varepsilon = \frac{1}{2(k-1)}$ and $t = C\sqrt{\sum_{i=1}^{k-1} p_i r_{i-1} r_i + p_k r_{k-1} + m}$, we have proved (49).

(3) For fixed $A \in \mathbb{R}^{r_k\times(p_1\cdots p_k)}, B \in \mathbb{R}^{(p_{k+1}\cdots p_d)\times(p_{k+1}r_{k+1})}$ satisfying $\|A\| \leq 1, \|B\| \leq 1$, by (47), we have

$$\begin{aligned}&\mathbb{P}\left(\|AZB\| \geq 2\sqrt{r_k + t}\right)\\ \leq& 2\cdot 5^{p_{k+1}r_{k+1}} \exp\left[-c\min\left(\frac{t^2}{r_k}, t\right)\right].\end{aligned}$$

Let

$$\begin{aligned}&M\\ =&\max_{(U_1,\ldots,V_d)\in\mathcal{A}} \left\|U_k^\top(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z\right.\\ &\left.\cdot(V_d\otimes I_{p_{k+1}\ldots p_{d-1}})\cdots(V_{k+2}\otimes I_{p_{k+1}})\right\|,\end{aligned}$$

By similar arguments as (59), one has

$$\begin{aligned}&\mathbb{P}\left(M \geq \frac{2\sqrt{r_k + t}}{1-(d-1)\varepsilon}\right)\\ \leq& 2((2+\varepsilon)/\varepsilon)^{\sum_{1\leq i\leq d, i\neq k+1} p_i r_{i-1} r_i} 5^{p_{k+1}r_{k+1}}\\ &\cdot\exp\left[-c\min\left(\frac{t^2}{r_k}, t\right)\right]\end{aligned}$$

for any $0 < \epsilon < \frac{1}{d}$. By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C\sum_{i=1}^d p_i r_{i-1} r_i$, we have proved the third part of Lemma 5.

(4) For fixed $U_1, \ldots, U_{d-1}$ satisfying $\|U_i\| \leq 1$, let $A = U_{d-1}^\top(I_{p_{d-1}}\otimes U_{d-2}^\top)\cdots(I_{p_2\cdots p_{d-1}}\otimes U_1^\top) \in \mathbb{R}^{r_{d-1}\times(p_1\cdots p_{d-1})}$, then $\|A\| \leq 1$. By (48) with $B = I_{p_d}$, we have

$$\mathbb{P}\left(\|AZ\|_{\mathrm{F}}^2 \geq p_d r_{d-1} + t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{p_d r_{d-1}}, t\right)\right].$$

Let

$$\begin{aligned}M =& \max_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i},\\ \|U_i\|\leq 1}} \|U_{d-1}^\top(I_{p_{d-1}}\otimes U_{d-2}^\top)\cdots\\ &(I_{p_2\cdots p_{d-1}}\otimes U_1^\top)Z\|_{\mathrm{F}}.\end{aligned}$$

The similar proof of (59) leads us to

$$\begin{aligned}&\mathbb{P}\left(M^2 \geq \frac{r_{d-1}p_d + t}{(1-\varepsilon(d-1))^2}\right)\\ \leq& 2\left((2+\varepsilon)/\varepsilon\right)^{\sum_{k=1}^{d-1} p_k r_{k-1} r_k} \exp\left[-c\min\left(\frac{t^2}{p_d r_{d-1}}, t\right)\right].\end{aligned} \tag{60}$$

for $0 < \varepsilon < \frac{1}{d-1}$. By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C\sum_{k=1}^d p_k r_{k-1} r_k$, we have arrived at (52).

(5) For fixed $A \in \mathbb{R}^{r_k\times(p_1\cdots p_k)}$, $B \in \mathbb{R}^{(p_{k+1}\cdots p_d)\times(p_{k+1}r_{k+1})}$, $\|A\| \leq 1, \|B\| \leq 1$, by (48), we have

$$\begin{aligned}&\mathbb{P}\left(\|AZB\|_{\mathrm{F}}^2 \geq p_{k+1}r_{k+1}r_k + t\right)\\ \leq& 2\exp\left[-c\min\left(\frac{t^2}{p_{k+1}r_{k+1}r_k}, t\right)\right].\end{aligned}$$

Let

$$M = \max_{(U_1,\ldots,V_d)\in\mathcal{A}} \left\| U_k^\top (I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top) Z \cdot (V_d \otimes I_{p_{k+1}\ldots p_{d-1}})\cdots(V_{k+2}\otimes I_{p_{k+1}}) \right\|_{\mathrm{F}},$$

Similarly to (59), for any $0<\varepsilon<\frac{1}{d-1}$, we have

$$\mathbb{P}\left(M \geq \frac{\sqrt{p_{k+1}r_{k+1}r_k + t}}{1-(d-1)\varepsilon}\right) \leq 2((2+\varepsilon)/\varepsilon)^{\sum_{1\leq i\leq d, i\neq k+1} p_i r_{i-1} r_i} \cdot \exp\left[-c\min\left(\frac{t^2}{p_{k+1}r_{k+1}r_k}, t\right)\right]. \tag{61}$$

By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C\sum_{i=1}^d p_i r_{i-1} r_i$, we have proved (53). □

Lemma 6: Suppose $X, Z \in \mathbb{R}^{p_1\times p_2}$, $\mathrm{rank}(X)=r$. Let $Y = X+Z$, $\widehat{U} = \mathrm{SVD}_r^L(Y)$, $\widehat{V} = \mathrm{SVD}_r^R(Y)$. Then we have

$$\max\{\|\widehat{U}_\perp^\top X\|, \|X\widehat{V}_\perp\|\} \leq 2\|Z\|,$$
$$\max\{\|\widehat{U}_\perp^\top X\|_{\mathrm{F}}, \|X\widehat{V}_\perp\|_{\mathrm{F}}\} \leq 2\min\{\|Z\|_{\mathrm{F}}, \sqrt{r}\|Z\|\}.$$

Proof of Lemma 6: See [48, Lemma 6] and [96, Theorem 1]. □

REFERENCES

[1] X. Bi, A. Qu, and X. Shen, "Multilayer tensor factorization with applications to recommender systems," *Ann. Statist.*, vol. 46, no. 6B, pp. 3308–3333, Dec. 2018.
[2] M. Nasiri, M. Rezghi, and B. Minaei, "Fuzzy dynamic tensor decomposition algorithm for recommender system," *UCT J. Res. Sci., Eng. Technol.*, vol. 2, no. 2, pp. 52–55, 2014.
[3] J. Wozniak *et al.*, "Neurocognitive and neuroimaging correlates of pediatric traumatic brain injury: A diffusion tensor imaging (DTI) study," *Arch. Clin. Neuropsychol.*, vol. 22, no. 5, pp. 555–568, Jun. 2007.
[4] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Amer. Stat. Assoc.*, vol. 108, no. 502, pp. 540–552, 2013.
[5] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.
[6] I. V. Oseledets and E. E. Tyrtyshnikov, "Breaking the curse of dimensionality, or how to use SVD in many dimensions," *SIAM J. Sci. Comput.*, vol. 31, no. 5, pp. 3744–3759, 2009.
[7] A. Cichocki *et al.*, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar. 2015.
[8] M. Mondelli and A. Montanari, "On the connection between learning two-layer neural networks and tensor decomposition," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1051–1060.
[9] K. Zhong, Z. Song, and I. S. Dhillon, "Learning non-overlapping convolutional neural networks with multiple kernels," 2017, *arXiv:1711.03440*.
[10] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 517–520.
[11] C. Zhang, R. Han, A. R. Zhang, and P. M. Voyles, "Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition," *Ultramicroscopy*, vol. 219, Dec. 2020, Art. no. 113123.
[12] A. Bhattacharya and D. B. Dunson, "Simplex factor models for multivariate unordered categorical data," *J. Amer. Stat. Assoc.*, vol. 107, no. 497, pp. 362–377, Mar. 2012.
[13] D. B. Dunson and C. Xing, "Nonparametric Bayes modeling of multivariate categorical data," *J. Amer. Stat. Assoc.*, vol. 104, no. 487, pp. 1042–1051, Sep. 2009.
[14] G. G. Calvi, A. Moniri, M. Mahfouz, Q. Zhao, and D. P. Mandic, "Tucker tensor layer in fully connected neural networks," 2019, *arXiv:1903.06133*.
[15] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 442–450.
[16] A. Novikov, A. Rodomanov, A. Osokin, and D. Vetrov, "Putting MRFs on a tensor train," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 811–819.
[17] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.
[18] M. Fannes, B. Nachtergaele, and R. F. Werner, "Finitely correlated states on quantum spin chains," *Commun. Math. Phys.*, vol. 144, no. 3, pp. 443–490, 1992.
[19] I. Oseledets, "A new tensor decomposition," *Doklady Math.*, vol. 80, no. 1, pp. 495–496, 2009.
[20] I. V. Oseledets and E. E. Tyrtyshnikov, "Recursive decomposition of multidimensional tensors," *Doklady Math.*, vol. 80, no. 1, pp. 460–462, Aug. 2009.
[21] R. Orús, "Tensor networks for complex quantum systems," *Nature Rev. Phys.*, vol. 1, no. 9, pp. 538–550, Sep. 2019.
[22] S. Bravyi, D. Gosset, and R. Movassagh, "Classical algorithms for quantum mean values," *Nature Phys.*, vol. 17, no. 3, pp. 337–341, Mar. 2021.
[23] M. Rakhuba and I. Oseledets, "Calculating vibrational spectra of molecules using tensor train decomposition," *J. Chem. Phys.*, vol. 145, no. 12, 2016, Art. no. 124101.
[24] U. Schollwöck, "The density-matrix renormalization group in the age of matrix product states," *Ann. Phys.*, vol. 326, no. 1, pp. 96–192, 2011.
[25] E. Stoudenmire and D. J. Schwab, "Supervised learning with tensor networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4799–4807.
[26] D. Bigoni, A. P. Engsig-Karup, and Y. M. Marzouk, "Spectral tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 38, no. 4, pp. A2405–A2439, Jan. 2016.
[27] I. Oseledets and E. Tyrtyshnikov, "TT-cross approximation for multidimensional arrays," *Linear Algebra Appl.*, vol. 432, no. 1, pp. 70–88, Jan. 2010.
[28] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, pp. 1–39, Nov. 2013.
[29] S. V. Dolgov and D. V. Savostyanov, "Alternating minimal energy methods for linear systems in higher dimensions," *SIAM J. Sci. Comput.*, vol. 36, no. 5, pp. A2248–A2271, Jan. 2014.
[30] Z. Song, D. P. Woodruff, and P. Zhong, "Relative error tensor low rank approximation," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2019, pp. 2772–2789.
[31] L. Li, W. Yu, and K. Batselier, "Faster tensor train decomposition for sparse data," *J. Comput. Appl. Math.*, vol. 405, May 2022, Art. no. 113972.
[32] C. Lubich, T. Rohwedder, R. Schneider, and B. Vandereycken, "Dynamical approximation by hierarchical tucker and tensor-train tensors," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 2, pp. 470–494, Jan. 2013.
[33] L. Grasedyck, M. Kluge, and S. Krämer, "Variants of alternating least squares tensor completion in the tensor train format," *SIAM J. Sci. Comput.*, vol. 37, no. 5, pp. A2424–A2450, Jan. 2015.
[34] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do, "Efficient tensor completion for color image and video recovery: Low-rank tensor train," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2466–2479, May 2017.
[35] M. M. Steinlechner, "Riemannian optimization for solving high-dimensional problems with low-rank tensor structure," EPFL, Lausanne, Switzerland, Tech. Rep., 2016.
[36] A. Novikov, P. Izmailov, V. Khrulkov, M. Figurnov, and I. Oseledets, "Tensor train decomposition on TensorFlow (T3F)," *J. Mach. Learn. Res.*, vol. 21, no. 30, pp. 1–7, 2020.
[37] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *Ann. Statist.*, vol. 46, no. 1, pp. 60–89, 2018.
[38] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko, "Unbiased risk estimates for singular value thresholding and spectral estimators," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4643–4657, Oct. 2013.
[39] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *Ann. Statist.*, vol. 42, no. 6, pp. 2413–2440, 2014.
[40] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[41] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Statist.*, vol. 43, no. 1, pp. 177–214, 2015.
[42] O. Klopp, "Matrix completion by singular value thresholding: Sharp bounds," *Electron. J. Statist.*, vol. 9, no. 2, pp. 2348–2369, Jan. 2015.

[43] H. Zhang, L. Z. Cheng, and W. Zhu, "A lower bound guaranteeing exact matrix completion via singular value thresholding algorithm," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 3, pp. 454–459, Nov. 2011.

[44] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Ann. Statist.*, vol. 36, no. 6, pp. 2791–2817, 2008.

[45] A. Zhang and M. Wang, "Spectral state compression of Markov processes," *IEEE Trans. Inf. Theory*, vol. 66, no. 5, pp. 3202–3231, May 2020.

[46] L. de Lathauwer, N. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, Jul. 2000.

[47] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(R_1, R_2,..., R_N) approximation of higher-order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, Jan. 2000.

[48] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338, Nov. 2018.

[49] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen, "A new truncation strategy for the higher-order singular value decomposition," *SIAM J. Sci. Comput.*, vol. 34, no. 2, pp. A1027–A1052, Jan. 2012.

[50] A. Zhang and R. Han, "Optimal sparse singular value decomposition for high-dimensional high-order data," *J. Amer. Stat. Assoc.*, vol. 114, no. 528, pp. 1708–1725, 2019.

[51] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[52] V. Sharan and G. Valiant, "Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3095–3104.

[53] S. E. Leurgans, R. T. Ross, and R. B. Abel, "A decomposition for three-way arrays," *SIAM J. Matrix Anal. Appl.*, vol. 14, no. 4, pp. 1064–1083, Oct. 1993.

[54] M. Rajih, P. Comon, and R. A. Harshman, "Enhanced line search: A novel method to accelerate PARAFAC," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1128–1147, 2008.

[55] N. Colombo and N. Vlassis, "Tensor decomposition via joint matrix Schur decomposition," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2820–2828.

[56] A. Anandkumar, Y. Deng, R. Ge, and H. Mobahi, "Homotopy analysis for tensor PCA," in *Proc. Conf. Learn. Theory*, 2017, pp. 79–104.

[57] G. B. Arous, S. Mei, A. Montanari, and M. Nica, "The landscape of the spiked tensor model," *Commun. Pure Appl. Math.*, vol. 72, no. 11, pp. 2282–2330, Nov. 2019.

[58] S. B. Hopkins, J. Shi, and D. Steurer, "Tensor principal component analysis via sum-of-square proofs," in *Proc. Conf. Learn. Theory*, 2015, pp. 956–1006.

[59] Y. Luo and A. R. Zhang, "Tensor clustering with planted structures: Statistical optimality and computational limits," 2020, *arXiv:2005.10743*.

[60] A. Perry, A. S. Wein, and A. S. Bandeira, "Statistical limits of spiked tensor models," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 56, no. 1, pp. 230–264, 2020.

[61] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2897–2905.

[62] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, "Statistical and computational phase transitions in spiked tensor estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 511–515.

[63] G. Allen, "Sparse higher-order principal components analysis," in *Proc. Artif. Intell. Statist.*, 2012, pp. 27–36.

[64] G. I. Allen, "Regularized tensor factorizations and higher-order principal components analysis," 2012, *arXiv:1202.2476*.

[65] Y. Liu, L. Chen, and C. Zhu, "Improved robust tensor principal component analysis via low-rank core matrix," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1378–1389, Dec. 2018.

[66] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5249–5257.

[67] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 925–938, Apr. 2020.

[68] P. Zhou and J. Feng, "Outlier-robust tensor PCA," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2263–2271.

[69] Y. Luo, G. Raskutti, M. Yuan, and A. R. Zhang, "A sharp blockwise tensor perturbation bound for orthogonal iteration," *J. Mach. Learn. Res.*, vol. 22, no. 179, pp. 1–48, 2021.

[70] A. S. Wein, A. El Alaoui, and C. Moore, "The Kikuchi hierarchy and tensor PCA," in *Proc. IEEE 60th Annu. Symp. Found. Comput. Sci. (FOCS)*, Nov. 2019, pp. 1446–1468.

[71] A. R. Benson, D. F. Gleich, and L.-H. Lim, "The spacey random walk: A stochastic process for higher-order data," *SIAM Rev.*, vol. 59, no. 2, pp. 321–345, 2017.

[72] A. E. Raftery, "A model for high-order Markov chains," *J. Roy. Stat. Soc. B, Methodol.*, vol. 47, no. 3, pp. 528–539, Jul. 1985.

[73] R. S. Tsay, *Analysis of Financial Time Series*, vol. 543. Hoboken, NJ, USA: Wiley, 2005.

[74] J. Zhao and S. Sun, "High-order Gaussian process dynamical models for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2014–2019, Jul. 2016.

[75] A. Berchtold and A. Raftery, "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series," *Stat. Sci.*, vol. 17, no. 3, pp. 328–356, Aug. 2002.

[76] A. Ganguly, T. Petrov, and H. Koeppl, "Markov chain aggregation and its applications to combinatorial reaction networks," *J. Math. Biol.*, vol. 69, no. 3, pp. 767–797, Sep. 2014.

[77] Z. Du, N. Ozay, and L. Balzano, "Mode clustering for Markov jump systems," in *Proc. IEEE 8th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2019, pp. 126–130.

[78] J. Sanders, A. Proutière, and S.-Y. Yun, "Clustering in block Markov chains," *Ann. Statist.*, vol. 48, no. 6, pp. 3488–3512, Dec. 2020.

[79] Z. Zhu, X. Li, M. Wang, and A. Zhang, "Learning Markov models via low-rank optimization," 2019, *arXiv:1907.00113*.

[80] M. J. Kearns and S. P. Singh, "Finite-sample convergence rates for Q-learning and indirect algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 996–1002.

[81] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l_1-ball for learning in high dimensions," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 272–279.

[82] R. Han, Y. Luo, M. Wang, and A. R. Zhang, "Exact clustering in tensor block model: Statistical optimality and computational limit," 2020, *arXiv:2012.09996*.

[83] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape Urban Planning*, vol. 106, no. 1, pp. 73–87, 2012.

[84] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. London, U.K.: Springer, 2009.

[85] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[86] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data," *Bioinformatics*, vol. 23, no. 12, pp. 1537–1544, 2007.

[87] D. S. Chaplot, P. Bhattacharyya, and A. Paranjape, "Unsupervised word sense disambiguation using Markov random field and dependency parser," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[88] M. I. Jordan and M. J. Wainwright, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, 2008.

[89] Y. Duan, M. Wang, Z. Wen, and Y. Yuan, "Adaptive low-nonnegative-rank approximation for state aggregation of Markov chains," *SIAM J. Matrix Anal. Appl.*, vol. 41, no. 1, pp. 244–278, Jan. 2020.

[90] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[91] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Reinforcement learning with soft state aggregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1995, pp. 361–368.

[92] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, vol. 135. Cambridge, MA, USA: MIT Press, 1998.

[93] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2010, *arXiv:1011.3027*.

[94] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *Ann. Statist.*, vol. 41, no. 6, pp. 3074–3110, 2013.

[95] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[96] Y. Luo, R. Han, and A. R. Zhang, "A Schatten-q low-rank matrix perturbation analysis via perturbation projection error bound," *Linear Algebra Appl.*, vol. 630, pp. 225–240, Dec. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0024379521002962

Yuchen Zhou received the B.E. degree from Peking University in 2016 and the Ph.D. degree in statistics from the University of Wisconsin–Madison in 2021. He is currently a Post-Doctoral Researcher with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania. His research interests include high-dimensional statistical inference, tensor data analysis, reinforcement learning, and statistical learning theory.

Lili Zheng received the bachelor's degree from the University of Science and Technology of China (USTC) in 2016 and the Ph.D. degree in statistics from the University of Wisconsin–Madison in 2021. She is currently a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Rice University. Her research interests span dependent data, high-dimensional statistics, network analysis, tensor modeling, stochastic algorithms, and non-convex optimization.

Anru R. Zhang received the bachelor's degree from Peking University in 2010 and the Ph.D. degree from the University of Pennsylvania in 2015. He is currently the Eugene Anson Stead, Jr., M.D. Associate Professor with the Department of Biostatistics and Bioinformatics and an Associate Professor with the Departments of Computer Science, Mathematics, and Statistical Science, Duke University. He was an Assistant Professor of statistics with the University of Wisconsin–Madison from 2015 to 2021. His work focuses on high-dimensional statistical inference, non-convex optimization, statistical tensor analysis, computational complexity, and applications in genomics, microbiome, electronic health records, and computational imaging. He received the ASA Gottfried E. Noether Junior Award (2021), the Bernoulli Society New Researcher Award in 2021, the ICSA Outstanding Young Researcher Award in 2021, and the NSF CAREER Award in 2020.

Yazhen Wang received the Ph.D. degree in statistics from the University of California at Berkeley in 1992. He is currently the Chair and a Professor of statistics with the University of Wisconsin–Madison. His research interests are financial econometrics, quantum computation, machine learning, high-dimensional statistics, nonparametric curve estimation, wavelets, change points, long-memory processes, and order restricted inference. He is a fellow of the Institute of Mathematical Statistics (IMS) and the American Statistical Association (ASA). He served on numerous professional committees of ASA, IMS, and ICSA, and the NSF Program Director. He was an Editor of *Statistica Sinica* and *Statistics and Its Interface*, and an Associate Editor of various journals, including *Annals of Statistics*, *Annals of Applied Statistics*, *Journal of the American Statistical Association*, *Journal of Business and Economic Statistics*, and *Statistica Sinica*.