Tensor Robust Principal Component Analysis with A New Tensor Nuclear Norm

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Member, IEEE, Zhouchen Lin, *Fellow, IEEE*, and Shuicheng Yan, *Fellow, IEEE*

Abstract—In this paper, we consider the Tensor Robust Principal Component Analysis (TRPCA) problem, which aims to exactly recover the low-rank and sparse components from their sum. Our model is based on the recently proposed tensor-tensor product (or t-product) [15]. Induced by the t-product, we first rigorously deduce the tensor spectral norm, tensor nuclear norm, and tensor average rank, and show that the tensor nuclear norm is the convex envelope of the tensor average rank within the unit ball of the tensor spectral norm. These definitions, their relationships and properties are consistent with matrix cases. Equipped with the new tensor nuclear norm, we then solve the TRPCA problem by solving a convex program and provide the theoretical guarantee for the exact recovery. Our TRPCA model and recovery guarantee include matrix RPCA as a special case. Numerical experiments verify our results, and the applications to image recovery and background modeling problems demonstrate the effectiveness of our method.

Index Terms—Tensor robust PCA, convex optimization, tensor nuclear norm, tensor singular value decomposition

1 Introduction

Principal Component Analysis (PCA) is a fundamental approach for data analysis. It exploits low-dimensional structure in high-dimensional data, which commonly exists in different types of data, *e.g.*, image, text, video and bioinformatics. It is computationally efficient and powerful for data instances which are mildly corrupted by small noises. However, a major issue of PCA is that it is brittle to be grossly corrupted or outlying observations, which are ubiquitous in real-world data. To date, a number of robust versions of PCA have been proposed, but many of them suffer from a high computational cost.

The Robust PCA [3] is the first polynomial-time algorithm with strong recovery guarantees. Suppose that we are given an observed matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, which can be decomposed as $\boldsymbol{X} = \boldsymbol{L}_0 + \boldsymbol{E}_0$, where $\boldsymbol{L}_0$ is low-rank and $\boldsymbol{E}_0$ is sparse. It is shown in [3] that if the singular vectors of $\boldsymbol{L}_0$ satisfy some incoherent conditions, *e.g.*, $\boldsymbol{L}_0$ is low-rank and $\boldsymbol{E}_0$ is sufficiently sparse, then $\boldsymbol{L}_0$ and $\boldsymbol{E}_0$ can be exactly recovered with high probability by solving the following convex problem

$$\min_{\boldsymbol{L},\boldsymbol{E}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{E}\|_1, \text{ s.t. } \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E}, \tag{1}$$

- *C. Lu is with the Department of Electrical and Computer Engineering, Carnegie Mellon University (e-mail: canyilu@gmail.com).*
- *J. Feng and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore (e-mail: elefjia@nus.edu.sg; eleyans@nus.edu.sg).*
- *Y. Chen is with the School of Operations Research and Information Engineering, Cornell University (e-mail: yudong.chen@cornell.edu).*
- *W. Liu is with the Tencent AI Lab, Shenzhen, China (e-mail: wl2223@columbia.edu).*
- *Z. Lin is with the Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: zlin@pku.edu.cn).*

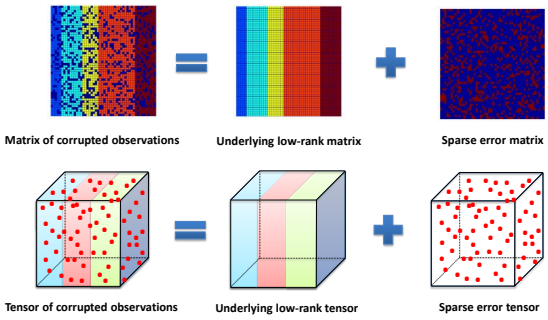


Fig. 1: Illustrations of RPCA [3] (up row) and our Tensor RPCA (bottom row). RPCA: low-rank and sparse matrix decomposition from noisy matrix observations. Tensor RPCA: low-rank and sparse tensor decomposition from noisy tensor observations.

where $\|\boldsymbol{L}\|_*$ denotes the nuclear norm (sum of the singular values of $\boldsymbol{L}$), and $\|\boldsymbol{E}\|_1$ denotes the ℓ_1-norm (sum of the absolute values of all the entries in $\boldsymbol{E}$). Theoretically, RPCA is guaranteed to work even if the rank of $\boldsymbol{L}_0$ grows almost linearly in the dimension of the matrix, and the errors in $\boldsymbol{E}_0$ are up to a constant fraction of all entries. The parameter λ is suggested to be set as $1/\sqrt{\max(n_1, n_2)}$ which works well in practice. Algorithmically, program (1) can be solved by efficient algorithms, at a cost not too much higher than PCA. RPCA and its extensions have been successfully applied to background modeling [3], subspace clustering [17], video compressive sensing [31], *etc*.

One major shortcoming of RPCA is that it can only handle 2-way (matrix) data. However, real data is usually multi-dimensional in nature-the information is stored in multi-way arrays known as tensors [16]. For example, a color image is a 3-way object with column, row and color modes; a greyscale video

is indexed by two spatial variables and one temporal variable. To use RPCA, one has to first restructure the multi-way data into a matrix. Such a preprocessing usually leads to an information loss and would cause a performance degradation. To alleviate this issue, it is natural to consider extending RPCA to manipulate the tensor data by taking advantage of its multi-dimensional structure.

In this work, we are interested in the Tensor Robust Principal Component (TRPCA) model which aims to exactly recover a low-rank tensor corrupted by sparse errors. See Figure 1 for an intuitive illustration. More specifically, suppose that we are given a data tensor $\boldsymbol{\mathcal{X}}$, and know that it can be decomposed as

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{E}}_0, \tag{2}$$

where $\boldsymbol{\mathcal{L}}_0$ is low-rank and $\boldsymbol{\mathcal{E}}_0$ is sparse, and both components are of arbitrary magnitudes. Note that we do not know the locations of the nonzero elements of $\boldsymbol{\mathcal{E}}_0$, not even how many there are. Now we consider a similar problem to RPCA. Can we recover the low-rank and sparse components exactly and efficiently from $\boldsymbol{\mathcal{X}}$? This is the problem of tensor RPCA studied in this work.

The tensor extension of RPCA is not easy since the numerical algebra of tensors is fraught with hardness results [11], [5], [8]. A main issue is that the tensor rank is not well defined with a tight convex relaxation. Several tensor rank definitions and their convex relaxations have been proposed but each has its limitation. For example, the CP rank [16], defined as the smallest number of rank one tensor decomposition, is generally NP-hard to compute. Also its convex relaxation is intractable. This makes the low CP rank tensor recovery challenging. The tractable Tucker rank [16] and its convex relaxation are more widely used. For a k-way tensor $\boldsymbol{\mathcal{X}}$, the Tucker rank is a vector defined as $\text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) := \Big(\text{rank}(\boldsymbol{X}^{\{1\}}), \text{rank}(\boldsymbol{X}^{\{2\}}), \cdots, \text{rank}(\boldsymbol{X}^{\{k\}})\Big)$, where $\boldsymbol{X}^{\{i\}}$ is the mode-i matricization of $\boldsymbol{\mathcal{X}}$ [16]. Motivated by the fact that the nuclear norm is the convex envelope of the matrix rank within the unit ball of the spectral norm, the Sum of Nuclear Norms (SNN) [18], defined as $\sum_i \|\boldsymbol{X}^{\{i\}}\|_*$, is used as a convex surrogate of $\sum_i \text{rank}(\boldsymbol{X}^{\{i\}})$. Then the work [24] considers the Low-Rank Tensor Completion (LRTC) model based on SNN:

$$\min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^{k} \lambda_i \|\boldsymbol{X}^{\{i\}}\|_*, \text{ s.t. } \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{M}}), \tag{3}$$

where $\lambda_i > 0$, and $\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}})$ denotes the projection of $\boldsymbol{\mathcal{X}}$ on the observed set $\boldsymbol{\Omega}$. The effectiveness of this approach for image processing has been well studied in [18], [28]. However, SNN is not the convex envelope of $\sum_i \text{rank}(\boldsymbol{X}^{\{i\}})$ [26]. Actually, the above model can be substantially suboptimal [24]: reliably recovering a k-way tensor of length n and Tucker rank $(r, r, \cdots, r)$ from Gaussian measurements requires $O(rn^{k-1})$ observations. In contrast, a certain (intractable) nonconvex formulation needs only $O(rK + nrK)$ observations. A better (but still suboptimal) convexification based on a more balanced matricization is proposed in [24]. The work [13] presents the recovery guarantee for the SNN based tensor RPCA model

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{E}}} \sum_{i=1}^{k} \lambda_i \|\boldsymbol{L}^{\{i\}}\|_* + \|\boldsymbol{\mathcal{E}}\|_1, \text{ s.t. } \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}}. \tag{4}$$

A robust tensor CP decomposition problem is studied in [6]. Though the recovery is guaranteed, the algorithm is nonconvex.

The limitations of existing works motivate us to consider an interesting problem: is it possible to define a new tensor nuclear norm such that it is a tight convex surrogate of certain tensor rank, and thus its resulting tensor RPCA enjoys a similar tight recovery guarantee to that of the matrix RPCA? This work will provide a positive answer to this question. Our solution is inspired by the recently proposed tensor-tensor product (t-product) [15] which is a generalization of the matrix-matrix product. It enjoys several similar properties to the matrix-matrix product. For example, based on t-product, any tensors have the tensor Singular Value Decomposition (t-SVD) and this motivates a new tensor rank, *i.e.*, tensor tubal rank [14]. To recover a tensor of low tubal rank, we propose a new tensor nuclear norm which is rigorously induced by the t-product. First, the tensor spectral norm can be induced by the operator norm when treating the t-product as an operator. Then the tensor nuclear norm is defined as the dual norm of the tensor spectral norm. We further propose the tensor average rank (which is closely related to the tensor tubal rank), and prove that its convex envelope is the tensor nuclear norm within the unit ball of the tensor spectral norm. It is interesting that this framework, including the new tensor concepts and their relationships, is consistent with the one for the matrix cases. Equipped with these new tools, we then study the TRPCA problem which aims to recover the low tubal rank component $\boldsymbol{\mathcal{L}}_0$ and sparse component $\boldsymbol{\mathcal{E}}_0$ from noisy observations $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{E}}_0 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ (this work focuses on the 3-way tensor) by convex optimization

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{E}}} \|\boldsymbol{\mathcal{L}}\|_* + \lambda \|\boldsymbol{\mathcal{E}}\|_1, \text{ s.t. } \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}}, \tag{5}$$

where $\|\boldsymbol{\mathcal{L}}\|_*$ is our new tensor nuclear norm (see the definition in Section 3). We prove that under certain incoherence conditions, the solution to (5) perfectly recovers the low-rank and the sparse components, provided of course that the tubal rank of $\boldsymbol{\mathcal{L}}_0$ is not too large, and that $\boldsymbol{\mathcal{E}}_0$ is reasonably sparse. A remarkable fact, like in RPCA, is that (5) has no tunning parameter either. Our analysis shows that $\lambda = 1/\sqrt{\max(n_1, n_2) n_3}$ guarantees the exact recovery when $\boldsymbol{\mathcal{L}}_0$ and $\boldsymbol{\mathcal{E}}_0$ satisfy certain assumptions. As a special case, if $\boldsymbol{\mathcal{X}}$ reduces to a matrix ($n_3 = 1$ in this case), all the new tensor concepts reduce to the matrix cases. Our TRPCA model (5) reduces to RPCA in (1), and also our recovery guarantee in Theorem 4.1 reduces to Theorem 1.1 in [3]. Another advantage of (5) is that it can be solved by polynomial-time algorithms.

The contributions of this work are summarized as follows:

1. Motivated by the t-product [15] which is a natural generalization of the matrix-matrix product, we rigorously deduce a new tensor nuclear norm and some other related tensor concepts, and they own the same relationship as the matrix cases. This is the foundation for the extensions of the models, optimization method and theoretical analyzing techniques from matrix cases to tensor cases.
2. Equipped with the tensor nuclear norm, we theoretically show that under certain incoherence conditions, the solution to the convex TRPCA model (5) perfectly recovers the underlying

low-rank component $\mathcal{L}_0$ and sparse component $\mathcal{E}_0$. RPCA [3] and its recovery guarantee fall into our special cases.

3. We give a new rigorous proof of t-SVD factorization and a more efficient way than [19] for solving TRPCA. We further perform several simulations to corroborate our theoretical results. Numerical experiments on images and videos also show the superiority of TRPCA over RPCA and SNN.

The rest of this paper is structured as follows. Section A gives some notations and preliminaries. Section 3 presents the way for defining the tensor nuclear norm induced by the t-product. Section 4 provides the recovery guarantee of TRPCA and the optimization details. Section 5 presents numerical experiments conducted on synthetic and real data. We conclude this work in Section 6.

2 Notations and Preliminaries

2.1 Notations

In this paper, we denote tensors by boldface Euler script letters, *e.g.*, $\boldsymbol{\mathcal{A}}$. Matrices are denoted by boldface capital letters, *e.g.*, $\boldsymbol{A}$; vectors are denoted by boldface lowercase letters, *e.g.*, $\boldsymbol{a}$, and scalars are denoted by lowercase letters, *e.g.*, a. We denote $\boldsymbol{I}_n$ as the $n \times n$ identity matrix. The fields of real numbers and complex numbers are denoted as $\mathbb{R}$ and $\mathbb{C}$, respectively. For a 3-way tensor $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$, we denote its (i,j,k)-th entry as $\boldsymbol{\mathcal{A}}_{ijk}$ or a_{ijk} and use the Matlab notation $\boldsymbol{\mathcal{A}}(i,:,:)$, $\boldsymbol{\mathcal{A}}(:,i,:)$ and $\boldsymbol{\mathcal{A}}(:,:,i)$ to denote respectively the i-th horizontal, lateral and frontal slice (see definitions in [16]). More often, the frontal slice $\boldsymbol{\mathcal{A}}(:,:,i)$ is denoted compactly as $\boldsymbol{A}^{(i)}$. The tube is denoted as $\boldsymbol{\mathcal{A}}(i,j,:)$. The inner product between $\boldsymbol{A}$ and $\boldsymbol{B}$ in $\mathbb{C}^{n_1 \times n_2}$ is defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathtt{Tr}(\boldsymbol{A}^* \boldsymbol{B})$, where $\boldsymbol{A}^*$ denotes the conjugate transpose of $\boldsymbol{A}$ and $\mathtt{Tr}(\cdot)$ denotes the matrix trace. The inner product between $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ in $\mathbb{C}^{n_1 \times n_2 \times n_3}$ is defined as $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \sum_{i=1}^{n_3} \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{B}^{(i)} \right\rangle$. For any $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$, the complex conjugate of $\boldsymbol{\mathcal{A}}$ is denoted as $\mathtt{conj}(\boldsymbol{\mathcal{A}})$ which takes the complex conjugate of each entry of $\boldsymbol{\mathcal{A}}$. We denote $\lfloor t \rfloor$ as the nearest integer less than or equal to t and $\lceil t \rceil$ as the one greater than or equal to t.

Some norms of vector, matrix and tensor are used. We denote the ℓ_1-norm as $\|\boldsymbol{\mathcal{A}}\|_1 = \sum_{ijk} |a_{ijk}|$, the infinity norm as $\|\boldsymbol{\mathcal{A}}\|_\infty = \max_{ijk} |a_{ijk}|$ and the Frobenius norm as $\|\boldsymbol{\mathcal{A}}\|_F = \sqrt{\sum_{ijk} |a_{ijk}|^2}$, respectively. The above norms reduce to the vector or matrix norms if $\boldsymbol{\mathcal{A}}$ is a vector or a matrix. For $\boldsymbol{v} \in \mathbb{C}^n$, the ℓ_2-norm is $\|\boldsymbol{v}\|_2 = \sqrt{\sum_i |v_i|^2}$. The spectral norm of a matrix $\boldsymbol{A}$ is denoted as $\|\boldsymbol{A}\| = \max_i \sigma_i(\boldsymbol{A})$, where $\sigma_i(\boldsymbol{A})$'s are the singular values of $\boldsymbol{A}$. The matrix nuclear norm is $\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A})$.

2.2 Discrete Fourier Transformation

The Discrete Fourier Transformation (DFT) plays a core role in tensor-tensor product introduced later. We give some related background knowledge and notations here. The DFT on $\boldsymbol{v} \in \mathbb{R}^n$, denoted as $\bar{\boldsymbol{v}}$, is given by

$$\bar{\boldsymbol{v}} = \boldsymbol{F}_n \boldsymbol{v} \in \mathbb{C}^n, \tag{6}$$

where $\boldsymbol{F}_n$ is the DFT matrix defined as

$$\boldsymbol{F}_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)(n-1)} \end{bmatrix} \in \mathbb{C}^{n \times n},$$

where $\omega = \mathrm{e}^{-\frac{2\pi i}{n}}$ is a primitive n-th root of unity in which $i = \sqrt{-1}$. Note that $\boldsymbol{F}_n / \sqrt{n}$ is a unitary matrix, *i.e.*,

$$\boldsymbol{F}_n^* \boldsymbol{F}_n = \boldsymbol{F}_n \boldsymbol{F}_n^* = n \boldsymbol{I}_n. \tag{7}$$

Thus $\boldsymbol{F}_n^{-1} = \boldsymbol{F}_n^* / n$. The above property will be frequently used in this paper. Computing $\bar{\boldsymbol{v}}$ by using (6) costs $O(n^2)$. A more widely used method is the Fast Fourier Transform (FFT) which costs $O(n \log n)$. By using the Matlab command `fft`, we have $\bar{\boldsymbol{v}} = \mathtt{fft}(\boldsymbol{v})$. Denote the circulant matrix of $\boldsymbol{v}$ as

$$\mathtt{circ}(\boldsymbol{v}) = \begin{bmatrix} v_1 & v_n & \cdots & v_2 \\ v_2 & v_1 & \cdots & v_3 \\ \vdots & \vdots & \ddots & \vdots \\ v_n & v_{n-1} & \cdots & v_1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

It is known that it can be diagonalized by the DFT matrix, *i.e.*,

$$\boldsymbol{F}_n \cdot \mathtt{circ}(\boldsymbol{v}) \cdot \boldsymbol{F}_n^{-1} = \mathtt{Diag}(\bar{\boldsymbol{v}}), \tag{8}$$

where $\mathtt{Diag}(\bar{\boldsymbol{v}})$ denotes a diagonal matrix with its i-th diagonal entry as $\bar{v}_i$. The above equation implies that the columns of $\boldsymbol{F}_n$ are the eigenvectors of $(\mathtt{circ}(\boldsymbol{v}))^\top$ and $\bar{v}_i$'s are the corresponding eigenvalues.

Lemma 2.1. *[25] Given any real vector $\boldsymbol{v} \in \mathbb{R}^n$, the associated $\bar{\boldsymbol{v}}$ satisfies*

$$\bar{v}_1 \in \mathbb{R} \text{ and } \mathtt{conj}(\bar{v}_i) = \bar{v}_{n-i+2},\ i = 2, \cdots, \left\lfloor \frac{n+1}{2} \right\rfloor. \tag{9}$$

Conversely, for any given complex $\bar{\boldsymbol{v}} \in \mathbb{C}^n$ satisfying (9), there exists a real block circulant matrix $\mathtt{circ}(\boldsymbol{v})$ such that (8) holds.

As will be seen later, the above properties are useful for efficient computation and important for proofs. Now we consider the DFT on tensors. For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote $\bar{\boldsymbol{\mathcal{A}}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ as the result of DFT on $\boldsymbol{\mathcal{A}}$ along the 3-rd dimension, *i.e.*, performing the DFT on all the tubes of $\boldsymbol{\mathcal{A}}$. By using the Matlab command `fft`, we have

$$\bar{\boldsymbol{\mathcal{A}}} = \mathtt{fft}(\boldsymbol{\mathcal{A}}, [\,], 3).$$

In a similar fashion, we can compute $\boldsymbol{\mathcal{A}}$ from $\bar{\boldsymbol{\mathcal{A}}}$ using the inverse FFT, *i.e.*,

$$\boldsymbol{\mathcal{A}} = \mathtt{ifft}(\bar{\boldsymbol{\mathcal{A}}}, [\,], 3).$$

In particular, we denote $\bar{\boldsymbol{A}} \in \mathbb{C}^{n_1 n_3 \times n_2 n_3}$ as a block diagonal matrix with its i-th block on the diagonal as the i-th frontal slice $\bar{\boldsymbol{A}}^{(i)}$ of $\bar{\boldsymbol{\mathcal{A}}}$, *i.e.*,

$$\bar{\boldsymbol{A}} = \mathtt{bdiag}(\bar{\boldsymbol{\mathcal{A}}}) = \begin{bmatrix} \bar{\boldsymbol{A}}^{(1)} & & & \\ & \bar{\boldsymbol{A}}^{(2)} & & \\ & & \ddots & \\ & & & \bar{\boldsymbol{A}}^{(n_3)} \end{bmatrix},$$

where bdiag is an operator which maps the tensor $\bar{\mathcal{A}}$ to the block diagonal matrix $\bar{A}$. Also, we define the block circulant matrix $\texttt{bcirc}(\mathcal{A}) \in \mathbb{R}^{n_1 n_3 \times n_2 n_3}$ of $\mathcal{A}$ as

$$\texttt{bcirc}(\mathcal{A}) = \begin{bmatrix} \boldsymbol{A}^{(1)} & \boldsymbol{A}^{(n_3)} & \cdots & \boldsymbol{A}^{(2)} \\ \boldsymbol{A}^{(2)} & \boldsymbol{A}^{(1)} & \cdots & \boldsymbol{A}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}^{(n_3)} & \boldsymbol{A}^{(n_3-1)} & \cdots & \boldsymbol{A}^{(1)} \end{bmatrix}.$$

Just like the circulant matrix which can be diagonalized by DFT, the block circulant matrix can be block diagonalized, *i.e.*,

$$(\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1}) \cdot \texttt{bcirc}(\mathcal{A}) \cdot (\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_2}) = \bar{\boldsymbol{A}}, \tag{10}$$

where $\otimes$ denotes the Kronecker product and $(\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1})/\sqrt{n_3}$ is unitary. By using Lemma 2.1, we have

$$\begin{cases} \bar{\boldsymbol{A}}^{(1)} \in \mathbb{R}^{n_1 \times n_2}, \\ \texttt{conj}(\bar{\boldsymbol{A}}^{(i)}) = \bar{\boldsymbol{A}}^{(n_3-i+2)},\ i = 2, \cdots, \lfloor \frac{n_3+1}{2} \rfloor. \end{cases} \tag{11}$$

Conversely, for any given $\bar{\mathcal{A}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ satisfying (11), there exists a real tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that (10) holds. Also, by using (7), we have the following properties which will be used frequently:

$$\|\mathcal{A}\|_F = \frac{1}{\sqrt{n_3}} \|\bar{\boldsymbol{A}}\|_F, \tag{12}$$

$$\langle \mathcal{A}, \mathcal{B} \rangle = \frac{1}{n_3} \langle \bar{\boldsymbol{A}}, \bar{\boldsymbol{B}} \rangle. \tag{13}$$

2.3 T-product and T-SVD

For $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we define

$$\texttt{unfold}(\mathcal{A}) = \begin{bmatrix} \boldsymbol{A}^{(1)} \\ \boldsymbol{A}^{(2)} \\ \vdots \\ \boldsymbol{A}^{(n_3)} \end{bmatrix}, \ \texttt{fold}(\texttt{unfold}(\mathcal{A})) = \mathcal{A},$$

where the unfold operator maps $\mathcal{A}$ to a matrix of size $n_1 n_3 \times n_2$ and fold is its inverse operator.

Definition 2.1. ***(T-product)*** *[15] Let $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2 \times l \times n_3}$. Then the t-product $\mathcal{A} * \mathcal{B}$ is defined to be a tensor of size $n_1 \times l \times n_3$,*

$$\mathcal{A} * \mathcal{B} = \texttt{fold}(\texttt{bcirc}(\mathcal{A}) \cdot \texttt{unfold}(\mathcal{B})). \tag{14}$$

The t-product can be understood from two perspectives. First, in the original domain, a 3-way tensor of size $n_1 \times n_2 \times n_3$ can be regarded as an $n_1 \times n_2$ matrix with each entry being a tube that lies in the third dimension. Thus, the t-product is analogous to the matrix multiplication except that the circular convolution replaces the multiplication operation between the elements. Note that the t-product reduces to the standard matrix multiplication when $n_3 = 1$. This is a key observation which makes our tensor RPCA model shown later involve the matrix RPCA as a special case. Second, the t-product is equivalent to the matrix multiplication in the Fourier domain; that is, $\mathcal{C} = \mathcal{A} * \mathcal{B}$ is equivalent to $\bar{\boldsymbol{C}} = \bar{\boldsymbol{A}}\bar{\boldsymbol{B}}$ due to (10). Indeed, $\mathcal{C} = \mathcal{A} * \mathcal{B}$ implies

$$\begin{aligned} &\texttt{unfold}(\mathcal{C}) \\ =&\texttt{bcirc}(\mathcal{A}) \cdot \texttt{unfold}(\mathcal{B}) \\ =&(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_1}) \cdot ((\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1}) \cdot \texttt{bcirc}(\mathcal{A}) \cdot (\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_2})) \\ &\cdot ((\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_2}) \cdot \texttt{unfold}(\mathcal{B})) \\ =&(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_1}) \cdot \bar{\boldsymbol{A}} \cdot \texttt{unfold}(\bar{\boldsymbol{B}}), \end{aligned} \tag{15}$$

where (15) uses (10). Left multiplying both sides with $(\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1})$ leads to $\texttt{unfold}(\bar{\mathcal{C}}) = \bar{\boldsymbol{A}} \cdot \texttt{unfold}(\bar{\mathcal{B}})$. This is equivalent to $\bar{\boldsymbol{C}} = \bar{\boldsymbol{A}}\bar{\boldsymbol{B}}$. This property suggests an efficient way based on FFT to compute t-product instead of using (14). See Algorithm 1.

Algorithm 1 Tensor-Tensor Product

Input: $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $\mathcal{B} \in \mathbb{R}^{n_2 \times l \times n_3}$.
Output: $\mathcal{C} = \mathcal{A} * \mathcal{B} \in \mathbb{R}^{n_1 \times l \times n_3}$.

1. Compute $\bar{\mathcal{A}} = \texttt{fft}(\mathcal{A}, [\,], 3)$ and $\bar{\mathcal{B}} = \texttt{fft}(\mathcal{B}, [\,], 3)$.
2. Compute each frontal slice of $\bar{\mathcal{C}}$ by

$$\bar{\boldsymbol{C}}^{(i)} = \begin{cases} \bar{\boldsymbol{A}}^{(i)} \bar{\boldsymbol{B}}^{(i)}, & i = 1, \cdots, \lceil \frac{n_3+1}{2} \rceil, \\ \texttt{conj}(\bar{\boldsymbol{C}}^{(n_3-i+2)}), & i = \lceil \frac{n_3+1}{2} \rceil + 1, \cdots, n_3. \end{cases}$$

3. Compute $\mathcal{C} = \texttt{ifft}(\bar{\mathcal{C}}, [\,], 3)$.

The t-product enjoys many similar properties to the matrix-matrix product. For example, the t-product is associative, *i.e.*, $\mathcal{A} * (\mathcal{B} * \mathcal{C}) = (\mathcal{A} * \mathcal{B}) * \mathcal{C}$. We also need some other concepts on tensors extended from the matrix cases.

Definition 2.2. ***(Conjugate transpose)*** *The conjugate transpose of a tensor $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ is the tensor $\mathcal{A}^* \in \mathbb{C}^{n_2 \times n_1 \times n_3}$ obtained by conjugate transposing each of the frontal slices and then reversing the order of transposed frontal slices 2 through n_3.*

The tensor conjugate transpose extends the tensor transpose [15] for complex tensors. As an example, let $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times 4}$ and its frontal slices be $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$ and $\boldsymbol{A}_4$. Then

$$\mathcal{A}^* = \texttt{fold}\left(\begin{bmatrix} \boldsymbol{A}_1^* \\ \boldsymbol{A}_4^* \\ \boldsymbol{A}_3^* \\ \boldsymbol{A}_2^* \end{bmatrix} \right).$$

Definition 2.3. ***(Identity tensor)*** *[15] The identity tensor $\mathcal{I} \in \mathbb{R}^{n \times n \times n_3}$ is the tensor with its first frontal slice being the $n \times n$ identity matrix, and other frontal slices being all zeros.*

It is clear that $\mathcal{A} * \mathcal{I} = \mathcal{A}$ and $\mathcal{I} * \mathcal{A} = \mathcal{A}$ given the appropriate dimensions. The tensor $\bar{\mathcal{I}} = \texttt{fft}(\mathcal{I}, [\,], 3)$ is a tensor with each frontal slice being the identity matrix.

Definition 2.4. ***(Orthogonal tensor)*** *[15] A tensor $\mathcal{Q} \in \mathbb{R}^{n \times n \times n_3}$ is orthogonal if it satisfies $\mathcal{Q}^* * \mathcal{Q} = \mathcal{Q} * \mathcal{Q}^* = \mathcal{I}$.*

Definition 2.5. ***(F-diagonal Tensor)*** *[15] A tensor is called f-diagonal if each of its frontal slices is a diagonal matrix.*

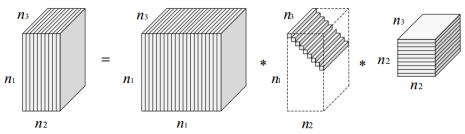


Fig. 2: An illustration of the t-SVD of an $n_1 \times n_2 \times n_3$ tensor [10].

Theorem 2.2. ***(T-SVD)*** *Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. Then it can be factorized as*

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*, \tag{16}$$

where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ are orthogonal, and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an f-diagonal tensor.

Proof. The proof is by construction. Recall that (10) holds and $\bar{\boldsymbol{A}}^{(i)}$'s satisfy the property (11). Then we construct the SVD of each $\bar{\boldsymbol{A}}^{(i)}$ in the following way. For $i = 1, \cdots, \lceil \frac{n_3+1}{2} \rceil$, let $\bar{\boldsymbol{A}}^{(i)} = \bar{\boldsymbol{U}}^{(i)} \bar{\boldsymbol{S}}^{(i)} (\bar{\boldsymbol{V}}^{(i)})^*$ be the full SVD of $\bar{\boldsymbol{A}}^{(i)}$. Here the singular values in $\bar{\boldsymbol{S}}^{(i)}$ are real. For $i = \lceil \frac{n_3+1}{2} \rceil + 1, \cdots, n_3$, let $\bar{\boldsymbol{U}}^{(i)} = \texttt{conj}(\bar{\boldsymbol{U}}^{(n_3-i+2)})$, $\bar{\boldsymbol{S}}^{(i)} = \bar{\boldsymbol{S}}^{(n_3-i+2)}$ and $\bar{\boldsymbol{V}}^{(i)} = \texttt{conj}(\bar{\boldsymbol{V}}^{(n_3-i+2)})$. Then, it is easy to verify that $\bar{\boldsymbol{A}}^{(i)} = \bar{\boldsymbol{U}}^{(i)} \bar{\boldsymbol{S}}^{(i)} (\bar{\boldsymbol{V}}^{(i)})^*$ gives the full SVD of $\bar{\boldsymbol{A}}^{(i)}$ for $i = \lceil \frac{n_3+1}{2} \rceil + 1, \cdots, n_3$. Then,

$$\bar{\boldsymbol{A}} = \bar{\boldsymbol{U}} \bar{\boldsymbol{S}} \bar{\boldsymbol{V}}^*. \tag{17}$$

By the construction of $\bar{\boldsymbol{U}}$, $\bar{\boldsymbol{S}}$ and $\bar{\boldsymbol{V}}$, and Lemma 2.1, we have that $(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_1}) \cdot \bar{\boldsymbol{U}} \cdot (\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1})$, $(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_1}) \cdot \bar{\boldsymbol{S}} \cdot (\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_2})$ and $(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_2}) \cdot \bar{\boldsymbol{V}} \cdot (\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_2})$ are real block circulant matrices. Then we can obtain an expression for $\texttt{bcirc}(\boldsymbol{\mathcal{A}})$ by applying the appropriate matrix $(\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_1})$ to the left and the appropriate matrix $(\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_2})$ to the right of each of the matrices in (17), and folding up the result. This gives a decomposition of the form $\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$, where $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{V}}$ are real. □

Theorem 2.2 shows that any 3 way tensor can be factorized into 3 components, including 2 orthogonal tensors and an f-diagonal tensor. See Figure 2 for an intuitive illustration of the t-SVD factorization. T-SVD reduces to the matrix SVD when $n_3 = 1$. We would like to emphasize that the result of Theorem 2.2 was given first in [15] and later in some other related works [10], [22]. But their proof and the way for computing $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ are not rigorous. The issue is that their method cannot guarantee that $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ are real tensors. They construct each frontal slice $\bar{\boldsymbol{U}}^{(i)}$ (or $\bar{\boldsymbol{V}}^{(i)}$) of $\bar{\boldsymbol{\mathcal{U}}}$ (or $\bar{\boldsymbol{\mathcal{V}}}$ resp.) from the SVD of $\bar{\boldsymbol{A}}^{(i)}$ independently for all $i = 1, \cdots, n_3$. However, the matrix SVD is not unique. Thus, $\bar{\boldsymbol{U}}^{(i)}$'s and $\bar{\boldsymbol{V}}^{(i)}$'s may not satisfy property (11) even though $\bar{\boldsymbol{A}}^{(i)}$'s do. In this case, the obtained $\boldsymbol{\mathcal{U}}$ (or $\boldsymbol{\mathcal{V}}$) from the inverse DFT of $\bar{\boldsymbol{\mathcal{U}}}$ (or $\bar{\boldsymbol{\mathcal{V}}}$ resp.) may not be real. Our proof above instead uses property (11) to construct $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ and thus avoids this issue. Our proof further leads to a more efficient way for computing t-SVD shown in Algorithm 2.

It is known that the singular values of a matrix have the decreasing order property. Let $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ be the t-SVD of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. The entries on the diagonal of the first frontal slice $\boldsymbol{\mathcal{S}}(:,:,1)$ of $\boldsymbol{\mathcal{S}}$ have the same decreasing property, *i.e.*,

$$\boldsymbol{\mathcal{S}}(1,1,1) \geq \boldsymbol{\mathcal{S}}(2,2,1) \geq \cdots \geq \boldsymbol{\mathcal{S}}(n',n',1) \geq 0, \tag{18}$$

Algorithm 2 T-SVD

Input: $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

Output: T-SVD components $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{V}}$ of $\boldsymbol{\mathcal{A}}$.

1. Compute $\bar{\boldsymbol{\mathcal{A}}} = \texttt{fft}(\boldsymbol{\mathcal{A}}, [\,], 3)$.
2. Compute each frontal slice of $\bar{\boldsymbol{\mathcal{U}}}$, $\bar{\boldsymbol{\mathcal{S}}}$ and $\bar{\boldsymbol{\mathcal{V}}}$ from $\bar{\boldsymbol{\mathcal{A}}}$ by
   for $i = 1, \cdots, \lceil \frac{n_3+1}{2} \rceil$ **do**
   $[\bar{\boldsymbol{U}}^{(i)}, \bar{\boldsymbol{S}}^{(i)}, \bar{\boldsymbol{V}}^{(i)}] = \text{SVD}(\bar{\boldsymbol{A}}^{(i)})$;
   end for
   for $i = \lceil \frac{n_3+1}{2} \rceil + 1, \cdots, n_3$ **do**
   $\bar{\boldsymbol{U}}^{(i)} = \texttt{conj}(\bar{\boldsymbol{U}}^{(n_3-i+2)})$;
   $\bar{\boldsymbol{S}}^{(i)} = \bar{\boldsymbol{S}}^{(n_3-i+2)}$;
   $\bar{\boldsymbol{V}}^{(i)} = \texttt{conj}(\bar{\boldsymbol{V}}^{(n_3-i+2)})$;
   end for
3. Compute $\boldsymbol{\mathcal{U}} = \texttt{ifft}(\bar{\boldsymbol{\mathcal{U}}}, [\,], 3)$, $\boldsymbol{\mathcal{S}} = \texttt{ifft}(\bar{\boldsymbol{\mathcal{S}}}, [\,], 3)$, and $\boldsymbol{\mathcal{V}} = \texttt{ifft}(\bar{\boldsymbol{\mathcal{V}}}, [\,], 3)$.

where $n' = \min(n_1, n_2)$. The above property holds since the inverse DFT gives

$$\boldsymbol{\mathcal{S}}(i,i,1) = \frac{1}{n_3} \sum_{j=1}^{n_3} \bar{\boldsymbol{\mathcal{S}}}(i,i,j), \tag{19}$$

and the entries on the diagonal of $\bar{\boldsymbol{\mathcal{S}}}(:,:,j)$ are the singular values of $\bar{\boldsymbol{\mathcal{A}}}(:,:,j)$. As will be seen in Section 3, the tensor nuclear norm depends only on the first frontal slice $\boldsymbol{\mathcal{S}}(:,:,1)$. Thus, we call the entries on the diagonal of $\boldsymbol{\mathcal{S}}(:,:,1)$ as the singular values of $\boldsymbol{\mathcal{A}}$.

Definition 2.6. ***(Tensor tubal rank)*** *[14], [34] For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the tensor tubal rank, denoted as $\text{rank}_t(\boldsymbol{\mathcal{A}})$, is defined as the number of nonzero singular tubes of $\boldsymbol{\mathcal{S}}$, where $\boldsymbol{\mathcal{S}}$ is from the t-SVD of $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$. We can write*

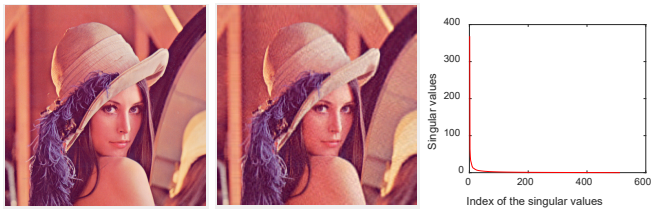
$$\text{rank}_t(\boldsymbol{\mathcal{A}}) = \#\{i, \boldsymbol{\mathcal{S}}(i,i,:) \neq \boldsymbol{0}\}.$$

By using property (19), the tensor tubal rank is determined by the first frontal slice $\boldsymbol{\mathcal{S}}(:,:,1)$ of $\boldsymbol{\mathcal{S}}$, *i.e.*,

$$\text{rank}_\text{t}(\boldsymbol{\mathcal{A}}) = \#\{i, \boldsymbol{\mathcal{S}}(i,i,1) \neq 0\}.$$

Hence, the tensor tubal rank is equivalent to the number of nonzero singular values of $\boldsymbol{\mathcal{A}}$. This property is the same as the matrix case. Define $\boldsymbol{\mathcal{A}}_k = \sum_{i=1}^{k} \boldsymbol{\mathcal{U}}(:,i,:) * \boldsymbol{\mathcal{S}}(i,i,:) * \boldsymbol{\mathcal{V}}(:,i,:)^*$ for some $k < \min(n_1, n_2)$. Then $\boldsymbol{\mathcal{A}}_k = \arg\min_{\text{rank}_t(\tilde{\boldsymbol{\mathcal{A}}}) \leq k} \|\boldsymbol{\mathcal{A}} - \tilde{\boldsymbol{\mathcal{A}}}\|_F$, so $\boldsymbol{\mathcal{A}}_k$ is the best approximation of $\boldsymbol{\mathcal{A}}$ with the tubal rank at most k. It is known that the real color images can be well approximated by low-rank matrices on the three channels independently. If we treat a color image as a three way tensor with each channel corresponding to a frontal slice, then it can be well approximated by a tensor of low tubal rank. A similar observation was found in [10] with the application to facial recognition. Figure 3 gives an example to show that a color image can be well approximated by a low tubal rank tensor since most of the singular values of the corresponding tensor are relatively small.

In Section 3, we will define a new tensor nuclear norm which is the convex surrogate of the tensor average rank defined as follows. This rank is closely related to the tensor tubal rank.



(a) (b) (c)

Fig. 3: Color images can be approximated by low tubal rank tensors. (a) A color image can be modeled as a tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{512\times512\times3}$; (b) approximation by a tensor with tubal rank $r = 50$; (c) plot of the singular values of $\boldsymbol{\mathcal{M}}$.

Definition 2.7. ***(Tensor average rank)*** *For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, the tensor average rank, denoted as $rank_a(\boldsymbol{\mathcal{A}})$, is defined as*

$$rank_a(\boldsymbol{\mathcal{A}}) = \frac{1}{n_3} rank(\texttt{bcirc}(\boldsymbol{\mathcal{A}})). \tag{20}$$

The above definition has a factor $\frac{1}{n_3}$. Note that this factor is crucial in this work as it guarantees that the convex envelope of the tensor average rank within a certain set is the tensor nuclear norm defined in Section 3. The underlying reason for this factor is the t-product definition. Each element of $\boldsymbol{\mathcal{A}}$ is repeated n_3 times in the block circulant matrix $\texttt{bcirc}(\boldsymbol{\mathcal{A}})$ used in the t-product. Intuitively, this factor alleviates such an entries expansion issue.

There are some connections between different tensor ranks and these properties imply that the low tubal rank or low average rank assumptions are reasonable for their applications in real visual data. First, $\text{rank}_\text{a}(\boldsymbol{\mathcal{A}}) \leq \text{rank}_\text{t}(\boldsymbol{\mathcal{A}})$. Indeed,

$$\text{rank}_\text{a}(\boldsymbol{\mathcal{A}}) = \frac{1}{n_3}\text{rank}(\bar{\boldsymbol{A}}) \leq \max_{i=1,\cdots,n_3} \text{rank}(\bar{\boldsymbol{A}}^{(i)}) = \text{rank}_\text{t}(\boldsymbol{\mathcal{A}}),$$

where the first equality uses (10). This implies that a low tubal rank tensor always has low average rank. Second, let $\text{rank}_\text{tc}(\boldsymbol{\mathcal{A}}) = \Big(\text{rank}(\boldsymbol{A}^{\{1\}}), \text{rank}(\boldsymbol{A}^{\{2\}}), \text{rank}(\boldsymbol{A}^{\{3\}})\Big)$, where $\boldsymbol{A}^{\{i\}}$ is the mode-i matricization of $\boldsymbol{\mathcal{A}}$, be the Tucker rank of $\boldsymbol{\mathcal{A}}$. Then $\text{rank}_\text{a}(\boldsymbol{\mathcal{A}}) \leq \text{rank}(\boldsymbol{A}^{\{1\}})$. This implies that a tensor with low Tucker rank has low average rank. The low Tucker rank assumption used in some applications, *e.g.*, image completion [18], is applicable to the low average rank assumption. Third, if the CP rank of $\boldsymbol{\mathcal{A}}$ is r, then its tubal rank is at most r [33]. Let $\boldsymbol{\mathcal{A}} = \sum_{i=1}^{r} \boldsymbol{a}_i^{(1)} \circ \boldsymbol{a}_i^{(2)} \circ \boldsymbol{a}_i^{(3)}$, where $\circ$ denotes the outer product, be the CP decomposition of $\boldsymbol{\mathcal{A}}$. Then $\bar{\boldsymbol{\mathcal{A}}} = \sum_{i=1}^{r} \boldsymbol{a}_i^{(1)} \circ \boldsymbol{a}_i^{(2)} \circ \bar{\boldsymbol{a}}_i^{(3)}$, where $\bar{\boldsymbol{a}}_i^{(3)} = \texttt{fft}(\boldsymbol{a}_i^{(3)})$. So $\bar{\boldsymbol{\mathcal{A}}}$ has the CP rank at most r, and each frontal slice of $\bar{\boldsymbol{\mathcal{A}}}$ is the sum of r rank-1 matrices. Thus, the tubal rank of $\boldsymbol{\mathcal{A}}$ is at most r. In summary, we show that the low average rank assumption is weaker than the low Tucker rank and low CP rank assumptions.

3 Tensor Nuclear Norm (TNN)

In this section, we propose a new tensor nuclear norm which is a convex surrogate of tensor average rank. Based on t-SVD, one may have many different ways to define the tensor nuclear norm intuitively. We give a new and rigorous way to deduce the tensor nuclear norm from the t-product, such that the concepts and their properties are consistent with the matrix cases. This is important since it guarantees that the theoretical analysis of the tensor nuclear norm based tensor RPCA model in Section 4 can be done in a similar way to RPCA. Figure 4 summarizes the way for the new definitions and their relationships. It begins with the known operator norm [1] and t-product. First, the tensor spectral norm is induced by the tensor operator norm by treating the t-product as an operator. Then the tensor nuclear norm is defined as the dual norm of the tensor spectral norm. Finally, we show that the tensor nuclear norm is the convex envelope of the tensor average rank within the unit ball of the tensor spectral norm.

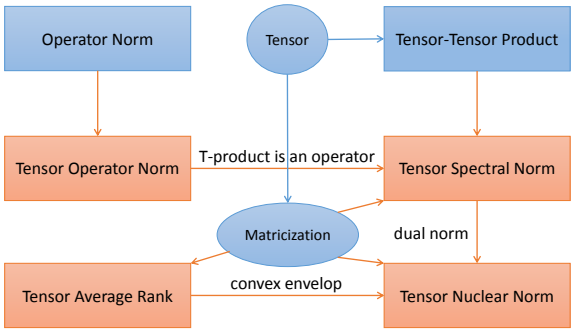


Fig. 4: An illustration of the way to define the tensor nuclear norm and the relationship with other tensor concepts. First, the tensor operator norm is a special case of the known operator norm performed on the tensors. The tensor spectral norm is induced by the tensor operator norm by treating the tensor-tensor product as an operator. Then the tensor nuclear norm is defined as the dual norm of the tensor spectral norm. We also define the tensor average rank and show that its convex envelope is the tensor nuclear norm within the unit ball of the tensor spectral norm. As detailed in Section 3, the tensor spectral norm, tensor nuclear norm and tensor average rank are also defined on the matricization of the tensor.

Let us first recall the concept of operator norm [1]. Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed linear spaces and $L : V \to W$ be the bounded linear operator between them, respectively. The operator norm of L is defined as

$$\|L\| = \sup_{\|\boldsymbol{v}\|_V \leq 1} \|L(\boldsymbol{v})\|_W. \tag{21}$$

Let $V = \mathbb{C}^{n_2}$, $W = \mathbb{C}^{n_1}$ and $L(\boldsymbol{v}) = \boldsymbol{A}\boldsymbol{v}$, $\boldsymbol{v} \in V$, where $\boldsymbol{A} \in \mathbb{C}^{n_1\times n_2}$. Based on different choices of $\|\cdot\|_V$ and $\|\cdot\|_W$, many matrix norms can be induced by the operator norm in (21). For example, if $\|\cdot\|_V$ and $\|\cdot\|_W$ are $\|\cdot\|_F$, then the operator norm (21) reduces to the matrix spectral norm.

Now, consider the normed linear spaces $(V, \|\cdot\|_F)$ and $(W, \|\cdot\|_F)$, where $V = \mathbb{R}^{n_2\times1\times n_3}$, $W = \mathbb{R}^{n_1\times1\times n_3}$, and $\boldsymbol{\mathcal{L}} : V \to W$ is a bounded linear operator. In this case, (21) reduces to the tensor operator norm

$$\|\boldsymbol{\mathcal{L}}\| = \sup_{\|\boldsymbol{\mathcal{V}}\|_F \leq 1} \|\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{V}})\|_F. \tag{22}$$

As a special case, if $\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{V}}) = \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{V}}$, where $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\boldsymbol{\mathcal{V}} \in V$, then the tensor operator norm (22) gives the tensor

spectral norm, denoted as $\|\boldsymbol{\mathcal{A}}\|$,

$$\begin{aligned}\|\boldsymbol{\mathcal{A}}\| :=& \sup_{\|\boldsymbol{\mathcal{V}}\|_F\leq 1} \|\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{V}}\|_F \\ =& \sup_{\|\boldsymbol{\mathcal{V}}\|_F\leq 1} \|\texttt{bcirc}(\boldsymbol{\mathcal{A}}) \cdot \texttt{unfold}(\boldsymbol{\mathcal{V}})\|_F \quad (23)\\ =&\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|, \quad (24)\end{aligned}$$

where (23) uses (14), and (24) uses the definition of matrix spectral norm.

Definition 3.1. ***(Tensor spectral norm)*** *The tensor spectral norm of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ is defined as $\|\boldsymbol{\mathcal{A}}\| := \|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|$.*

By (7) and (10), we have

$$\|\boldsymbol{\mathcal{A}}\| = \|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\| = \|\bar{\boldsymbol{A}}\|. \quad (25)$$

This property is frequently used in this work. It is known that the matrix nuclear norm is the dual norm of the matrix spectral norm. Thus, we define the tensor nuclear norm, denoted as $\|\boldsymbol{\mathcal{A}}\|_*$, as the dual norm of the tensor spectral norm. For any $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\tilde{\boldsymbol{B}} \in \mathbb{C}^{n_1n_3\times n_2n_3}$, we have

$$\begin{aligned}\|\boldsymbol{\mathcal{A}}\|_* :=& \sup_{\|\boldsymbol{\mathcal{B}}\|\leq 1} \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\rangle \quad (26)\\ =& \sup_{\|\bar{\boldsymbol{B}}\|\leq 1} \frac{1}{n_3}\langle \bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}\rangle \quad (27)\\ \leq& \frac{1}{n_3} \sup_{\|\tilde{\boldsymbol{B}}\|\leq 1} |\langle \bar{\boldsymbol{A}}, \tilde{\boldsymbol{B}}\rangle| \quad (28)\\ =& \frac{1}{n_3}\|\bar{\boldsymbol{A}}\|_*, \quad (29)\\ =& \frac{1}{n_3}\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|_*, \quad (30)\end{aligned}$$

where (27) is from (13), (28) is due to the fact that $\bar{\boldsymbol{B}}$ is a block diagonal matrix in $\mathbb{C}^{n_1n_3\times n_2n_3}$ while $\tilde{\boldsymbol{B}}$ is an arbitrary matrix in $\mathbb{C}^{n_1n_3\times n_2n_3}$, (29) uses the fact that the matrix nuclear norm is the dual norm of the matrix spectral norm, and (30) uses (10) and (7). Now we show that there exists $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ such that the equality (28) holds and thus $\|\boldsymbol{\mathcal{A}}\|_* = \frac{1}{n_3}\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|_*$. Let $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ be the t-SVD of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*$. We have

$$\begin{aligned}\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\rangle =& \langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*, \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\rangle \quad (31)\\ =& \frac{1}{n_3}\left\langle \overline{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*}, \overline{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*}\right\rangle \\ =& \frac{1}{n_3}\left\langle \bar{\boldsymbol{U}}\bar{\boldsymbol{S}}\bar{\boldsymbol{V}}^*, \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^*\right\rangle = \frac{1}{n_3}\texttt{Tr}(\bar{\boldsymbol{S}}) \\ =& \frac{1}{n_3}\|\bar{\boldsymbol{A}}\|_* = \frac{1}{n_3}\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|_*. \quad (32)\end{aligned}$$

Combining (26)-(30) and (31)-(32) leads to $\|\boldsymbol{\mathcal{A}}\|_* = \frac{1}{n_3}\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|_*$. On the other hand, by (31)-(32), we have

$$\begin{aligned}\|\boldsymbol{\mathcal{A}}\|_* =& \langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*, \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\rangle \\ =& \langle \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{V}}^* * \boldsymbol{\mathcal{V}}\rangle \\ =& \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{I}}\rangle = \sum_{i=1}^{r} \boldsymbol{\mathcal{S}}(i,i,1), \quad (33)\end{aligned}$$

where $r = \text{rank}_\text{t}(\boldsymbol{\mathcal{A}})$ is the tubal rank. Thus, we have the following definition of tensor nuclear norm.

Definition 3.2. ***(Tensor nuclear norm)*** *Let $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ be the t-SVD of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$. The tensor nuclear norm of $\boldsymbol{\mathcal{A}}$ is defined as*

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

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

From (33), it can be seen that only the information in the first frontal slice of $\boldsymbol{\mathcal{S}}$ is used when defining the tensor nuclear norm. Note that this is the first work which directly uses the singular values $\boldsymbol{\mathcal{S}}(:,:,1)$ of a tensor to define the tensor nuclear norm. Such a definition makes it consistent with the matrix nuclear norm. The above TNN definition is also different from existing works [19], [34], [27].

It is known that the matrix nuclear norm $\|\boldsymbol{A}\|_*$ is the convex envelope of the matrix rank $\text{rank}(\boldsymbol{A})$ within the set $\{\boldsymbol{A}|\|\boldsymbol{A}\| \leq 1\}$ [9]. Now we show that the tensor average rank and tensor nuclear norm have the same relationship.

Theorem 3.1. *On the set $\{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3} | \|\boldsymbol{\mathcal{A}}\| \leq 1\}$, the convex envelope of the tensor average rank $rank_a(\boldsymbol{\mathcal{A}})$ is the tensor nuclear norm $\|\boldsymbol{\mathcal{A}}\|_*$.*

We would like to emphasize that the proposed tensor spectral norm, tensor nuclear norm and tensor ranks are not arbitrarily defined. They are rigorously induced by the t-product and t-SVD. These concepts and their relationships are consistent with the matrix cases. This is important for the proofs, analysis and computation in optimization. Table 1 summarizes the parallel concepts in sparse vector, low-rank matrix and low-rank tensor. With these elements in place, the existing proofs of low-rank matrix recovery provide a template for the more general case of low-rank tensor recovery.

Also, from the above discussions, we have the property

$$\|\boldsymbol{\mathcal{A}}\|_* = \frac{1}{n_3}\|\texttt{bcirc}(\boldsymbol{\mathcal{A}})\|_* = \frac{1}{n_3}\|\bar{\boldsymbol{A}}\|_*. \quad (34)$$

It is interesting to understand the tensor nuclear norm from the perspectives of $\texttt{bcirc}(\boldsymbol{\mathcal{A}})$ and $\bar{\boldsymbol{A}}$. The block circulant matrix can be regarded as a new way of matricization of $\boldsymbol{\mathcal{A}}$ in the original domain. The frontal slices of $\boldsymbol{\mathcal{A}}$ are arranged in a circulant way, which is expected to preserve more spacial relationships across frontal slices, compared with previous matricizations along a single dimension. Also, the block diagonal matrix $\bar{\boldsymbol{A}}$ can be regarded as a matricization of $\boldsymbol{\mathcal{A}}$ in the Fourier domain. Its blocks on the diagonal are the frontal slices of $\bar{\boldsymbol{\mathcal{A}}}$, which contains the information across frontal slices of $\boldsymbol{\mathcal{A}}$ due to the DFT on $\boldsymbol{\mathcal{A}}$ along the third dimension. So $\texttt{bcirc}(\boldsymbol{\mathcal{A}})$ and $\bar{\boldsymbol{A}}$ play a similar role to matricizations of $\boldsymbol{\mathcal{A}}$ in different domains. Both of them capture the spacial information within and across frontal slices of $\boldsymbol{\mathcal{A}}$. This intuitively supports our tensor nuclear norm definition.

Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^*$ be the skinny SVD of $\boldsymbol{A}$. It is known that any subgradient of the nuclear norm at $\boldsymbol{A}$ is of the form $\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}$, where $\boldsymbol{U}^*\boldsymbol{W} = \boldsymbol{0}$, $\boldsymbol{W}\boldsymbol{V} = \boldsymbol{0}$ and $\|\boldsymbol{W}\| \leq 1$ [32].

TABLE 1: Parallelism of sparse vector, low-rank matrix and low-rank tensor.

	Sparse vector	Low-rank matrix	Low-rank tensor (this work)
Degeneracy of	1-D signal $\boldsymbol{x} \in \mathbb{R}^n$	2-D correlated signals $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$	3-D correlated signals $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$
Parsimony concept	cardinality	rank	tensor average rank[1]
Measure	ℓ_0-norm $\|\boldsymbol{x}\|_0$	$\text{rank}(\boldsymbol{X})$	$\text{rank}_\text{a}(\boldsymbol{\mathcal{X}})$
Convex surrogate	ℓ_1-norm $\|\boldsymbol{x}\|_1$	nuclear norm $\|\boldsymbol{X}\|_*$	tensor nuclear norm $\|\boldsymbol{\mathcal{X}}\|_*$
Dual norm	ℓ_∞-norm $\|\boldsymbol{x}\|_\infty$	spectral norm $\|\boldsymbol{X}\|$	tensor spectral norm $\|\boldsymbol{\mathcal{X}}\|$

[2]Strictly speaking, the tensor tubal rank, which bounds the tensor average rank, is also the parsimony concept of the low-rank tensor.

Similarly, for $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with tubal rank r, we also have the skinny t-SVD, *i.e.*, $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$, where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1 \times r \times n_3}$, $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r \times r \times n_3}$, and $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2 \times r \times n_3}$, in which $\boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{V}}^* * \boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{I}}$. The skinny t-SVD will be used throughout this paper. With skinny t-SVD, we introduce the subgradient of the tensor nuclear norm, which plays an important role in the proofs.

Theorem 3.2. ***(Subgradient of tensor nuclear norm)*** *Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with $\text{rank}_t(\boldsymbol{\mathcal{A}}) = r$ and its skinny t-SVD be $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$. The subdifferential (the set of subgradients) of $\|\boldsymbol{\mathcal{A}}\|_*$ is $\partial\|\boldsymbol{\mathcal{A}}\|_* = \{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}} | \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{W}} = \boldsymbol{0}, \boldsymbol{\mathcal{W}} * \boldsymbol{\mathcal{V}} = \boldsymbol{0}, \|\boldsymbol{\mathcal{W}}\| \leq 1\}$.*

4 Exact Recovery Guarantee of TRPCA

With TNN defined above, we now consider the exact recovery guarantee of TRPCA in (5). The problem we study here is to recover a low tubal rank tensor $\boldsymbol{\mathcal{L}}_0$ from highly corrupted measurements $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{S}}_0$. In this section, we show that under certain assumptions, the low tubal rank part $\boldsymbol{\mathcal{L}}_0$ and sparse part $\boldsymbol{\mathcal{S}}_0$ can be exactly recovered by solving convex program (5). We will also give the optimization detail for solving (5).

4.1 Tensor Incoherence Conditions

Recovering the low-rank and sparse components from their sum suffers from an identifiability issue. For example, the tensor $\boldsymbol{\mathcal{X}}$, with $x_{ijk} = 1$ when $i = j = k = 1$ and zeros everywhere else, is both low-rank and sparse. One is not able to identify the low-rank component and the sparse component in this case. To avoid such pathological situations, we need to assume that the low-rank component $\boldsymbol{\mathcal{L}}_0$ is not sparse. To this end, we assume $\boldsymbol{\mathcal{L}}_0$ to satisfy some incoherence conditions. We denote $\mathring{\boldsymbol{e}}_i$ as the tensor column basis, which is a tensor of size $n_1 \times 1 \times n_3$ with its $(i, 1, 1)$-th entry equaling 1 and the rest equaling 0 [33]. We also define the tensor tube basis $\dot{\boldsymbol{e}}_k$, which is a tensor of size $1 \times 1 \times n_3$ with its $(1, 1, k)$-th entry equaling 1 and the rest equaling 0.

Definition 4.1. ***(Tensor Incoherence Conditions)*** *For $\boldsymbol{\mathcal{L}}_0 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, assume that $\text{rank}_t(\boldsymbol{\mathcal{L}}_0) = r$ and it has the skinny t-SVD $\boldsymbol{\mathcal{L}}_0 = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$, where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1 \times r \times n_3}$ and $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2 \times r \times n_3}$ satisfy $\boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{V}}^* * \boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{I}}$, and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r \times r \times n_3}$ is an f-diagonal tensor. Then $\boldsymbol{\mathcal{L}}_0$ is said to satisfy the tensor incoherence conditions with parameter μ if*

$$\max_{i=1,\cdots,n_1} \|\boldsymbol{\mathcal{U}}^* * \mathring{\boldsymbol{e}}_i\|_F \leq \sqrt{\frac{\mu r}{n_1 n_3}}, \tag{35}$$

$$\max_{j=1,\cdots,n_2} \|\boldsymbol{\mathcal{V}}^* * \mathring{\boldsymbol{e}}_j\|_F \leq \sqrt{\frac{\mu r}{n_2 n_3}}, \tag{36}$$

$$\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2 n_3^2}}. \tag{37}$$

The exact recovery guarantee of RPCA [3] also requires some incoherence conditions. Due to property (12), conditions (48)-(49) have equivalent matrix forms in the Fourier domain, and they are intuitively similar to the matrix incoherence conditions (1.2) in [3]. But the joint incoherence condition (50) is more different from the matrix case (1.3) in [3], since it does not have an equivalent matrix form in the Fourier domain. As observed in [4], the joint incoherence condition is not necessary for low-rank matrix completion. However, for RPCA, it is unavoidable for polynomial-time algorithms. In our proofs, the joint incoherence (50) condition is necessary. Another identifiability issue arises if the sparse tensor $\boldsymbol{\mathcal{S}}_0$ has low tubal rank. This can be avoided by assuming that the support of $\boldsymbol{\mathcal{S}}_0$ is uniformly distributed.

4.2 Main Results

Now we show that the convex program (5) is able to perfectly recover the low-rank and sparse components. We define $n_{(1)} = \max(n_1, n_2)$ and $n_{(2)} = \min(n_1, n_2)$.

Theorem 4.1. *Suppose that $\boldsymbol{\mathcal{L}}_0 \in \mathbb{R}^{n \times n \times n_3}$ obeys (48)-(50). Fix any $n \times n \times n_3$ tensor $\boldsymbol{\mathcal{M}}$ of signs. Suppose that the support set $\boldsymbol{\Omega}$ of $\boldsymbol{\mathcal{S}}_0$ is uniformly distributed among all sets of cardinality m, and that $\text{sgn}([\boldsymbol{\mathcal{S}}_0]_{ijk}) = [\boldsymbol{\mathcal{M}}]_{ijk}$ for all $(i, j, k) \in \boldsymbol{\Omega}$. Then, there exist universal constants $c_1, c_2 > 0$ such that with probability at least $1 - c_1(nn_3)^{-c_2}$ (over the choice of support of $\boldsymbol{\mathcal{S}}_0$), $(\boldsymbol{\mathcal{L}}_0, \boldsymbol{\mathcal{S}}_0)$ is the unique minimizer to (5) with $\lambda = 1/\sqrt{nn_3}$, provided that*

$$\text{rank}_t(\boldsymbol{\mathcal{L}}_0) \leq \frac{\rho_r n n_3}{\mu(\log(nn_3))^2} \text{ and } m \leq \rho_s n^2 n_3, \tag{38}$$

where ρ_r and ρ_s are positive constants. If $\boldsymbol{\mathcal{L}}_0 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ has rectangular frontal slices, TRPCA with $\lambda = 1/\sqrt{n_{(1)} n_3}$ succeeds with probability at least $1 - c_1(n_{(1)} n_3)^{-c_2}$, provided that $\text{rank}_t(\boldsymbol{\mathcal{L}}_0) \leq \frac{\rho_r n_{(2)} n_3}{\mu(\log(n_{(1)} n_3))^2}$ and $m \leq \rho_s n_1 n_2 n_3$.

The above result shows that for incoherent $\boldsymbol{\mathcal{L}}_0$, the perfect recovery is guaranteed with high probability for $\text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0)$ on the order of $nn_3/(\mu(\log nn_3)^2)$ and a number of nonzero entries in $\boldsymbol{\mathcal{S}}_0$ on the order of $n^2 n_3$. For $\boldsymbol{\mathcal{S}}_0$, we make only one assumption on the random location distribution, but no assumption about the magnitudes or signs of the nonzero entries. Also TRPCA is parameter free. The mathematical analysis implies that the parameter $\lambda = 1/\sqrt{nn_3}$ leads to the correct recovery. Moreover, since the t-product of 3-way tensors reduces to the standard matrix-matrix product when $n_3 = 1$, the tensor nuclear norm reduces to the matrix nuclear norm. Thus, RPCA is a special case of TRPCA and the guarantee of RPCA in Theorem 1.1 in [3] is a special case of our Theorem 4.1. Both our model and theoretical

guarantee are consistent with RPCA. Compared with SNN [13], our tensor extension of RPCA is much more simple and elegant.

The detailed proof of Theorem 4.1 can be found in the supplementary material. It is interesting to understand our proof from the perspective of the following equivalent formulation

$$\min_{\boldsymbol{\mathcal{L}},\,\boldsymbol{\mathcal{E}}} \frac{1}{n_3}\left(\|\bar{\boldsymbol{L}}\|_* + \lambda\|\texttt{bcirc}(\boldsymbol{\mathcal{E}})\|_1\right),\ \text{s.t. } \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}}, \quad (39)$$

where (34) is used. Program (39) is a mixed model since the low-rank regularization is performed on the Fourier domain while the sparse regularization is performed on the original domain. Our proof of Theorem 4.1 is also conducted based on the interaction between both domains. By interpreting the tensor nuclear norm of $\boldsymbol{\mathcal{L}}$ as the matrix nuclear norm of $\bar{\boldsymbol{L}}$ (with a factor $\frac{1}{n_3}$) in the Fourier domain, we are then able to use some existing properties of the matrix nuclear norm in the proofs. The analysis for the sparse term is kept on the original domain since the ℓ_1-norm has no equivalent form in the Fourier domain. Though both two terms of the objective function of (39) are given on two matrices ($\bar{\boldsymbol{L}}$ and $\texttt{bcirc}(\boldsymbol{\mathcal{E}})$), the analysis for model (39) is very different from that of matrix RPCA. The matrices $\bar{\boldsymbol{L}}$ and $\texttt{bcirc}(\boldsymbol{\mathcal{E}})$ can be regarded as two matricizations of the tensor objects $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{E}}$, respectively. Their structures are more complicated than those in matrix RPCA, and thus make the proofs different from [3]. For example, our proofs require proving several bounds of norms on random tensors. Theses results and proofs, which have to use the properties of block circulant matrices and the Fourier transformation, are completely new. Some proofs are challenging due to the dependent structure of $\texttt{bcirc}(\boldsymbol{\mathcal{E}})$ for $\boldsymbol{\mathcal{E}}$ with an independent elements assumption. Also, TRPCA is of a different nature from the tensor completion problem [33]. The proof of the exact recovery of TRPCA is more challenging since the ℓ_1-norm (and its dual norm ℓ_∞-norm used in (50)) has no equivalent formulation in the Fourier domain.

It is worth mentioning that this work focuses on the analysis for 3-way tensors. But it is not difficult to generalize our model in (5) and results in Theorem 4.1 to the case of order-p ($p \geq 3$) tensors, by using the t-SVD for order-p tensors in [22].

When considering the application of TRPCA, the way for constructing a 3-way tensor from data is important. The reason is that the t-product is orientation dependent, and so is the tensor nuclear norm. Thus, the value of TNN may be different if the tensor is rotated. For example, a 3-channel color image can be formatted as 3 different sizes of tensors. Therefore, when using TRPCA which is based on TNN, one has to format the data into tensors in a proper way by leveraging some priori knowledge, *e.g.*, the low tubal rank property of the constructed tensor.

4.3 Tensor Singular Value Thresholding

Problem (5) can be solved by the standard Alternating Direction Method of Multiplier (ADMM) [20]. A key step is to compute the proximal operator of TNN

$$\min_{\boldsymbol{\mathcal{X}}\in\mathbb{R}^{n_1\times n_2\times n_3}} \tau\|\boldsymbol{\mathcal{X}}\|_* + \frac{1}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Y}}\|_F^2. \quad (40)$$

We show that it also has a closed-form solution as the proximal operator of the matrix nuclear norm. Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ be the tensor SVD of $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1\times n_2\times n_3}$. For each $\tau > 0$, we define the tensor Singular Value Thresholding (t-SVT) operator as follows

$$\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}}) = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}}_\tau * \boldsymbol{\mathcal{V}}^*, \quad (41)$$

where

$$\boldsymbol{\mathcal{S}}_\tau = \texttt{ifft}((\bar{\boldsymbol{\mathcal{S}}} - \tau)_+, [\,], 3). \quad (42)$$

Note that $\bar{\boldsymbol{\mathcal{S}}}$ is a real tensor. Above t_+ denotes the positive part of t, *i.e.*, $t_+ = \max(t, 0)$. That is, this operator simply applies a soft-thresholding rule to the singular values $\bar{\boldsymbol{\mathcal{S}}}$ (not $\boldsymbol{\mathcal{S}}$) of the frontal slices of $\bar{\boldsymbol{\mathcal{Y}}}$, effectively shrinking these towards zero. The t-SVT operator is the proximity operator associated with TNN.

Algorithm 3 Tensor Singular Value Thresholding (t-SVT)

Input: $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $\tau > 0$.
Output: $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ as defined in (41).

1. Compute $\bar{\boldsymbol{\mathcal{Y}}} = \texttt{fft}(\boldsymbol{\mathcal{Y}}, [\,], 3)$.
2. Perform matrix SVT on each frontal slice of $\bar{\boldsymbol{\mathcal{Y}}}$ by
   for $i = 1, \cdots, \lceil \frac{n_3+1}{2} \rceil$ **do**
   $[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = \text{SVD}(\bar{\boldsymbol{Y}}^{(i)})$;
   $\bar{\boldsymbol{W}}^{(i)} = \boldsymbol{U} \cdot (\boldsymbol{S} - \tau)_+ \cdot \boldsymbol{V}^*$;
   end for
   for $i = \lceil \frac{n_3+1}{2} \rceil + 1, \cdots, n_3$ **do**
   $\bar{\boldsymbol{W}}^{(i)} = \texttt{conj}(\bar{\boldsymbol{W}}^{(n_3-i+2)})$;
   end for
3. Compute $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}}) = \texttt{ifft}(\bar{\boldsymbol{\mathcal{W}}}, [\,], 3)$.

Theorem 4.2. *For any $\tau > 0$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, the tensor singular value thresholding operator (41) obeys*

$$\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}}) = \arg\min_{\boldsymbol{\mathcal{X}}\in\mathbb{R}^{n_1\times n_2\times n_3}} \tau\|\boldsymbol{\mathcal{X}}\|_* + \frac{1}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Y}}\|_F^2. \quad (43)$$

Proof. The required solution to (43) is a real tensor and thus we first show that $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ in (41) is real. Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ be the tensor SVD of $\boldsymbol{\mathcal{Y}}$. We know that the frontal slices of $\bar{\boldsymbol{\mathcal{S}}}$ satisfy the property (11) and so do the frontal slices of $(\bar{\boldsymbol{\mathcal{S}}} - \tau)_+$. By Lemma 2.1, $\boldsymbol{\mathcal{S}}_\tau$ in (42) is real. Thus, $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ in (41) is real. Secondly, by using properties (34) and (12), problem (43) is equivalent to

$$\begin{aligned}&\arg\min_{\boldsymbol{\mathcal{X}}} \frac{1}{n_3}(\tau\|\bar{\boldsymbol{X}}\|_* + \frac{1}{2}\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\|_F^2)\\ =&\arg\min_{\boldsymbol{\mathcal{X}}} \frac{1}{n_3}\sum_{i=1}^{n_3}(\tau\|\bar{\boldsymbol{X}}^{(i)}\|_* + \frac{1}{2}\|\bar{\boldsymbol{X}}^{(i)} - \bar{\boldsymbol{Y}}^{(i)}\|_F^2). \quad (44)\end{aligned}$$

By Theorem 2.1 in [2], we know that the i-th frontal slice of $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ solves the i-th subproblem of (44). Hence, $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ solves problem (43). □

Theorem 4.2 gives the closed-form of the t-SVT operator $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$, which is a natural extension of the matrix SVT [2]. Note that $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$ is real when $\boldsymbol{\mathcal{Y}}$ is real. By using property (11), Algorithm 3 gives an efficient way for computing $\mathcal{D}_\tau(\boldsymbol{\mathcal{Y}})$.

With t-SVT, we now give the details of ADMM to solve (5). The augmented Lagrangian function of (5) is

$$\begin{aligned}L(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{Y}}, \mu) =& \|\boldsymbol{\mathcal{L}}\|_* + \lambda\|\boldsymbol{\mathcal{E}}\|_1 + \langle \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{X}}\rangle\\ &+ \frac{\mu}{2}\|\boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{X}}\|_F^2.\end{aligned}$$

Algorithm 4 Solve (5) by ADMM

Initialize: $\boldsymbol{\mathcal{L}}_0 = \boldsymbol{\mathcal{S}}_0 = \boldsymbol{\mathcal{Y}}_0 = 0$, $\rho = 1.1$, $\mu_0 = 1\text{e}{-3}$, $\mu_{\max} = 1\text{e}10$, $\epsilon = 1\text{e}{-8}$.

while not converged **do**

1. Update $\boldsymbol{\mathcal{L}}_{k+1}$ by

$$\boldsymbol{\mathcal{L}}_{k+1} = \underset{\boldsymbol{\mathcal{L}}}{\operatorname{argmin}} \|\boldsymbol{\mathcal{L}}\|_* + \frac{\mu_k}{2}\left\|\boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{E}}_k - \boldsymbol{\mathcal{X}} + \frac{\boldsymbol{\mathcal{Y}}_k}{\mu_k}\right\|_F^2;$$

2. Update $\boldsymbol{\mathcal{E}}_{k+1}$ by

$$\boldsymbol{\mathcal{E}}_{k+1} = \underset{\boldsymbol{\mathcal{E}}}{\operatorname{argmin}}\ \lambda\|\boldsymbol{\mathcal{E}}\|_1 + \frac{\mu_k}{2}\left\|\boldsymbol{\mathcal{L}}_{k+1} + \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{X}} + \frac{\boldsymbol{\mathcal{Y}}_k}{\mu_k}\right\|_F^2;$$

3. $\boldsymbol{\mathcal{Y}}_{k+1} = \boldsymbol{\mathcal{Y}}_k + \mu_k(\boldsymbol{\mathcal{L}}_{k+1} + \boldsymbol{\mathcal{E}}_{k+1} - \boldsymbol{\mathcal{X}})$;
4. Update μ_{k+1} by $\mu_{k+1} = \min(\rho\mu_k, \mu_{\max})$;
5. Check the convergence conditions

$$\|\boldsymbol{\mathcal{L}}_{k+1} - \boldsymbol{\mathcal{L}}_k\|_\infty \le \epsilon,\ \|\boldsymbol{\mathcal{E}}_{k+1} - \boldsymbol{\mathcal{E}}_k\|_\infty \le \epsilon,$$
$$\|\boldsymbol{\mathcal{L}}_{k+1} + \boldsymbol{\mathcal{E}}_{k+1} - \boldsymbol{\mathcal{X}}\|_\infty \le \epsilon.$$

end while

Then $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{E}}$ can be updated by minimizing the augmented Lagrangian function L alternately. Both subproblems have closed-form solutions. See Algorithm 4 for the whole procedure. The main per-iteration cost lies in the update of $\boldsymbol{\mathcal{L}}_{k+1}$, which requires computing FFT and $\lceil \frac{n_3+1}{2} \rceil$ SVDs of $n_1 \times n_2$ matrices. The per-iteration complexity is $O\left(n_1 n_2 n_3 \log n_3 + n_{(1)} n_{(2)}^2 n_3\right)$.

5 Experiments

In this section, we conduct numerical experiments to verify our main results in Theorem 4.1. We first investigate the ability of the convex TRPCA model (5) to recover tensors with varying tubal rank and different levels of sparse noises. We then apply it for image recovery and background modeling. As suggested by Theorem 4.1, we set $\lambda = 1/\sqrt{n_{(1)} n_3}$ in all the experiments[2]. But note that it is possible to further improve the performance by tuning λ more carefully. The suggested value in theory provides a good guide in practice. All the simulations are conducted on a PC with an Intel Xeon E3-1270 3.60GHz CPU and 64GB memory.

5.1 Exact Recovery from Varying Fractions of Error

We first verify the correct recovery guarantee of Theorem 4.1 on randomly generated problems. We simply consider the tensors of size $n \times n \times n$, with varying dimension n =100 and 200. We generate a tensor with tubal rank r as a product $\boldsymbol{\mathcal{L}}_0 = \boldsymbol{\mathcal{P}} * \boldsymbol{\mathcal{Q}}^*$, where $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{Q}}$ are $n \times r \times n$ tensors with entries independently sampled from $\mathcal{N}(0, 1/n)$ distribution. The support set $\boldsymbol{\Omega}$ (with size m) of $\boldsymbol{\mathcal{E}}_0$ is chosen uniformly at random. For all $(i, j, k) \in \boldsymbol{\Omega}$, let $[\boldsymbol{\mathcal{E}}_0]_{ijk} = \boldsymbol{\mathcal{M}}_{ijk}$, where $\boldsymbol{\mathcal{M}}$ is a tensor with independent Bernoulli ± 1 entries.

Table 2 reports the recovery results based on varying choices of the tubal rank r of $\boldsymbol{\mathcal{L}}_0$ and the sparsity m of $\boldsymbol{\mathcal{E}}_0$. It can be seen that our convex program (5) gives the correct tubal rank estimation of $\boldsymbol{\mathcal{L}}_0$ in all cases and also the relative errors $\|\hat{\boldsymbol{\mathcal{L}}} - \boldsymbol{\mathcal{L}}_0\|_F / \|\boldsymbol{\mathcal{L}}_0\|_F$ are very small, less than 10^{-5}. The sparsity estimation of $\boldsymbol{\mathcal{E}}_0$ is not as exact as the rank estimation, but note that the relative errors $\|\hat{\boldsymbol{\mathcal{E}}} - \boldsymbol{\mathcal{E}}_0\|_F / \|\boldsymbol{\mathcal{E}}_0\|_F$ are all very small, less than 10^{-8} (actually much smaller than the relative errors of the recovered low-rank component). These results well verify the correct recovery phenomenon as claimed in Theorem 4.1.

2. Codes of our method available at https://github.com/canyilu.

TABLE 2: Correct recovery for random problems of varying sizes.

$r = \text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0) = 0.05n, m = \|\boldsymbol{\mathcal{E}}_0\|_0 = 0.05n^3$

n	r	m	$\text{rank}_\text{t}(\hat{\boldsymbol{\mathcal{L}}})$	$\|\hat{\boldsymbol{\mathcal{S}}}\|_0$	$\frac{\|\hat{\boldsymbol{\mathcal{L}}}-\boldsymbol{\mathcal{L}}_0\|_F}{\|\boldsymbol{\mathcal{L}}_0\|_F}$	$\frac{\|\hat{\boldsymbol{\mathcal{E}}}-\boldsymbol{\mathcal{E}}_0\|_F}{\|\boldsymbol{\mathcal{E}}_0\|_F}$
100	5	5e4	5	50,029	2.6e−7	5.4e−10
200	10	4e5	10	400,234	5.9e−7	6.7e−10

$r = \text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0) = 0.05n, m = \|\boldsymbol{\mathcal{E}}_0\|_0 = 0.1n^3$

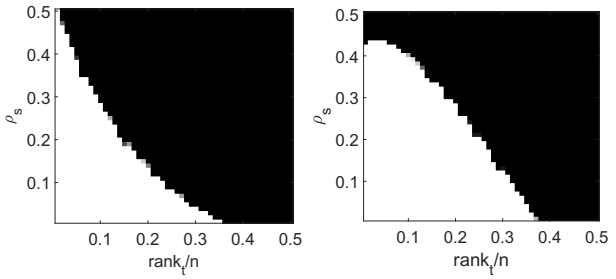
n	r	m	$\text{rank}_\text{t}(\hat{\boldsymbol{\mathcal{L}}})$	$\|\hat{\boldsymbol{\mathcal{S}}}\|_0$	$\frac{\|\hat{\boldsymbol{\mathcal{L}}}-\boldsymbol{\mathcal{L}}_0\|_F}{\|\boldsymbol{\mathcal{L}}_0\|_F}$	$\frac{\|\hat{\boldsymbol{\mathcal{E}}}-\boldsymbol{\mathcal{E}}_0\|_F}{\|\boldsymbol{\mathcal{E}}_0\|_F}$
100	5	1e5	5	100,117	4.1e−7	8.2e−10
200	10	8e5	10	800,901	4.4e−7	4.5e−10

$r = \text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0) = 0.1n, m = \|\boldsymbol{\mathcal{E}}_0\|_0 = 0.1n^3$

n	r	m	$\text{rank}_\text{t}(\hat{\boldsymbol{\mathcal{L}}})$	$\|\hat{\boldsymbol{\mathcal{S}}}\|_0$	$\frac{\|\hat{\boldsymbol{\mathcal{L}}}-\boldsymbol{\mathcal{L}}_0\|_F}{\|\boldsymbol{\mathcal{L}}_0\|_F}$	$\frac{\|\hat{\boldsymbol{\mathcal{E}}}-\boldsymbol{\mathcal{E}}_0\|_F}{\|\boldsymbol{\mathcal{E}}_0\|_F}$
100	10	1e5	10	101,952	4.8e−7	1.8e−9
200	20	8e5	20	815,804	4.9e−7	9.3e−10

$r = \text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0) = 0.1n, m = \|\boldsymbol{\mathcal{E}}_0\|_0 = 0.2n^3$

n	r	m	$\text{rank}_\text{t}(\hat{\boldsymbol{\mathcal{L}}})$	$\|\hat{\boldsymbol{\mathcal{E}}}\|_0$	$\frac{\|\hat{\boldsymbol{\mathcal{L}}}-\boldsymbol{\mathcal{L}}_0\|_F}{\|\boldsymbol{\mathcal{L}}_0\|_F}$	$\frac{\|\hat{\boldsymbol{\mathcal{E}}}-\boldsymbol{\mathcal{E}}_0\|_F}{\|\boldsymbol{\mathcal{E}}_0\|_F}$
100	10	2e5	10	200,056	7.7e−7	4.1e−9
200	20	16e5	20	1,601,008	1.2e−6	3.1e−9



(a) TRPCA, Random Signs (b) TRPCA, Coherent Signs

Fig. 5: Correct recovery for varying tubal ranks of $\boldsymbol{\mathcal{L}}_0$ and sparsities of $\boldsymbol{\mathcal{E}}_0$. Fraction of correct recoveries across 10 trials, as a function of $\text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0)$ (x-axis) and sparsity of $\boldsymbol{\mathcal{E}}_0$ (y-axis). Left: $\text{sgn}(\boldsymbol{\mathcal{E}}_0)$ random. Right: $\boldsymbol{\mathcal{E}}_0 = \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\text{sgn}(\boldsymbol{\mathcal{L}}_0)$.

5.2 Phase Transition in Tubal Rank and Sparsity

The results in Theorem 4.1 show the perfect recovery for incoherent tensor with $\text{rank}_\text{t}(\boldsymbol{\mathcal{L}}_0)$ on the order of $nn_3/(\mu(\log nn_3)^2)$ and the sparsity of $\boldsymbol{\mathcal{E}}_0$ on the order of $n^2 n_3$. Now we examine the recovery phenomenon with varying tubal rank of $\boldsymbol{\mathcal{L}}_0$ from varying sparsity of $\boldsymbol{\mathcal{E}}_0$. We consider the tensor $\boldsymbol{\mathcal{L}}_0$ of size $\mathbb{R}^{n\times n\times n_3}$, where $n = 100$ and $n_3 = 50$. We generate $\boldsymbol{\mathcal{L}}_0 = \boldsymbol{\mathcal{P}} * \boldsymbol{\mathcal{Q}}^*$, where $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{Q}}$ are $n \times r \times n_3$ tensors with entries independently sampled from a $\mathcal{N}(0, 1/n)$ distribution. For the sparse component $\boldsymbol{\mathcal{E}}_0$, we consider two cases. In the first case, we assume a Bernoulli model for the support of the sparse term $\boldsymbol{\mathcal{E}}_0$, with random signs: each entry of $\boldsymbol{\mathcal{E}}_0$ takes on value 0 with probability $1-\rho$, and values ± 1 each with probability $\rho/2$. The second case chooses the support $\boldsymbol{\Omega}$ in accordance with the Bernoulli model, but this time sets $\boldsymbol{\mathcal{E}}_0 = \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\text{sgn}(\boldsymbol{\mathcal{L}}_0)$. We set $\frac{r}{n} = [0.01 : 0.01 :$

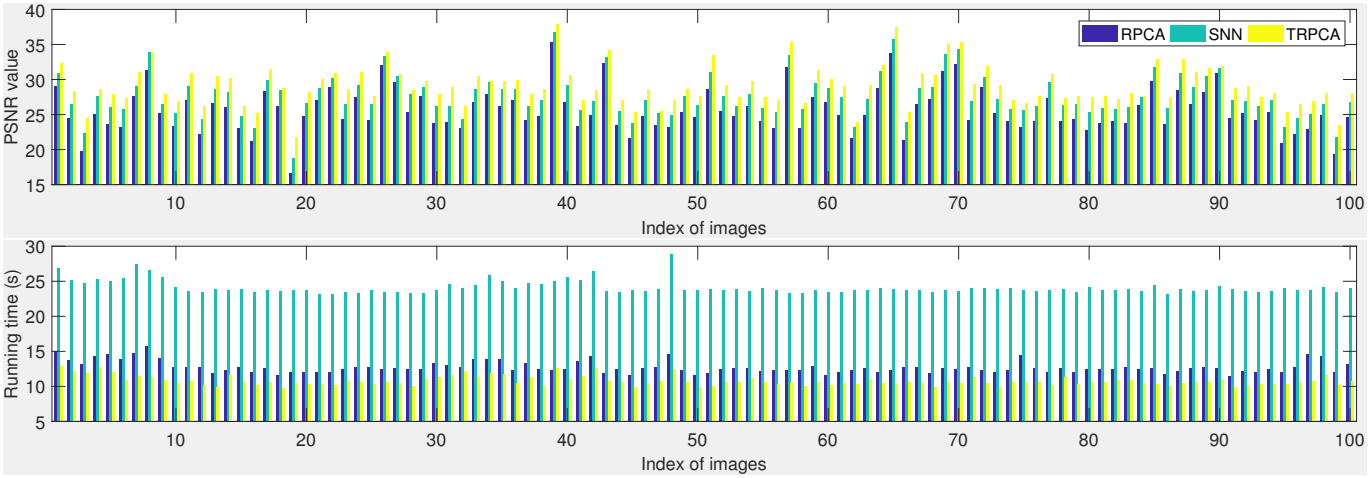


Fig. 6: Comparison of the PSNR values (top) and running time (bottom) obtained by RPCA, SNN and TRPCA on 100 images.

$0.5]$ and $\rho_s = [0.01 : 0.01 : 0.5]$. For each $(\frac{r}{n}, \rho_s)$-pair, we simulate 10 test instances and declare a trial to be successful if the recovered $\hat{\boldsymbol{\mathcal{L}}}$ satisfies $\|\hat{\boldsymbol{\mathcal{L}}} - \boldsymbol{\mathcal{L}}_0\|_F / \|\boldsymbol{\mathcal{L}}_0\|_F \leq 10^{-3}$. Figure 5 plots the fraction of correct recovery for each pair $(\frac{r}{n}, \rho_s)$ (black = 0% and white = 100%). It can be seen that there is a large region in which the recovery is correct in both cases. Intuitively, the experiment shows that the recovery is correct when the tubal rank of $\boldsymbol{\mathcal{L}}_0$ is relatively low and the errors $\boldsymbol{\mathcal{E}}_0$ is relatively sparse. Figure 5 (b) further shows that the signs of $\boldsymbol{\mathcal{E}}_0$ are not important: recovery can be guaranteed as long as its support is chosen uniformly at random. These observations are consistent with Theorem 4.1. Similar observations can be found in the matrix RPCA case (see Figure 1 in [3]).

5.3 Application to Image Recovery

We apply TRPCA to image recovery from the corrupted images with random noises.The motivation is that the color images can be approximated by low rank matrices or tensors [18]. We will show that the recovery performance of TRPCA is still satisfactory with the suggested parameter in theory on real data.

We use 100 color images from the Berkeley Segmentation Dataset [23] for the test. The sizes of images are 321×481 or 481×321. For each image, we randomly set 10% of pixels to random values in [0, 255], and the positions of the corrupted pixels are unknown. All the 3 channels of the images are corrupted at the same positions (the corruptions are on the whole tubes). This problem is more challenging than the corruptions on 3 channels at different positions. See Figure 7 (b) for some sample images with noises. We compare our TRPCA model with RPCA [3] and SNN [13] which also own the theoretical recovery guarantee. For RPCA, we apply it on each channel separably and combine the results to obtain the recovered image. The parameter λ is set to $\lambda = 1/\sqrt{\max(n_1, n_2)}$ as suggested in theory. For SNN in (4), we find that it does not perform well when λ_i's are set to the values suggested in theory [13]. We empirically set $\boldsymbol{\lambda} = [15, 15, 1.5]$ in (4) to make SNN perform well in most cases. For our TRPCA, we format a $n_1 \times n_2$ sized image as a tensor of size $n_1 \times n_2 \times 3$. We find that such a way of tensor construction usually performs better than some other ways. This may be due to the noises which present on the tubes. We set $\lambda = 1/\sqrt{3\max(n_1, n_2)}$ in TRPCA. We use the Peak Signal-to-Noise Ratio (PSNR), defined as

$$\text{PSNR} = 10\log_{10}\left(\frac{\|\boldsymbol{\mathcal{M}}\|_\infty^2}{\frac{1}{n_1 n_2 n_3}\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{M}}\|_F^2}\right),$$

to evaluate the recovery performance.

Figure 6 gives the comparison of the PSNR values and running time on all 100 images. Some examples with the recovered images are shown in Figure 7. From these results, we have the following observations. First, both SNN and TRPCA perform much better than the matrix based RPCA. The reason is that RPCA performs on each channel independently, and thus is not able to use the information across channels. The tensor methods instead take advantage of the multi-dimensional structure of data. Second, TRPCA outperforms SNN in most cases. This not only demonstrates the superiority of our TRPCA, but also validates our recovery guarantee in Theorem 4.1 on image data. Note that SNN needs some additional effort to tune the weighted parameters λ_i's empirically. Different from SNN which is a loose convex surrogate of the sum of Tucker rank, our TNN is a tight convex relaxation of the tensor average rank, and the recovery performance of the obtained optimal solutions has the tight recovery guarantee as RPCA. Third, we use the standard ADMM to solve RPCA, SNN and TRPCA. Figure 6 (bottom) shows that TRPCA is as efficient as RPCA, while SNN requires the highest cost in this experiment.

5.4 Application to Background Modeling

In this section, we consider the background modeling problem which aims to separate the foreground objects from the background. The frames of the background are highly correlated and thus can be modeled as a low rank tensor. The moving foreground objects occupy only a fraction of image pixels and thus can be treated as sparse errors. We solve this problem by using RPCA, SNN and TRPCA. We consider four color videos, *Hall*

(a) Orignal image (b) Observed image (c) RPCA (d) SNN (e) TRPCA

Index	1	2	3	4	5	6
RPCA	29.10	24.53	25.12	24.31	27.50	26.77
SNN	30.91	26.45	27.66	26.45	29.26	28.19
TRPCA	**32.33**	**28.30**	**28.59**	**28.62**	**31.06**	**30.16**

(f) Comparison of the PSNR values on the above 6 images.

Index	1	2	3	4	5	6
RPCA	14.98	13.79	14.35	12.45	12.72	15.73
SNN	26.93	25.20	25.33	23.47	23.38	28.16
TRPCA	**12.96**	**12.24**	**12.76**	**10.70**	**10.64**	**14.31**

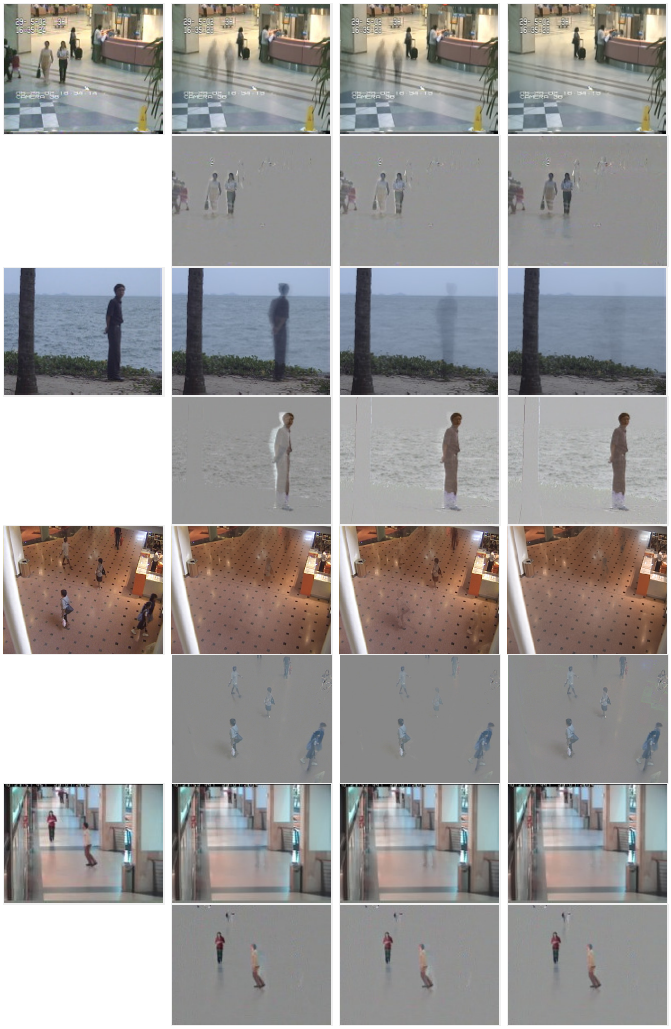
(g) Comparison of the running time (s) on the above 6 images.

Fig. 7: Recovery performance comparison on 6 example images. (a) Original image; (b) observed image; (c)-(e) recovered images by RPCA, SNN and TRPCA, respectively; (f) and (g) show the comparison of PSNR values and running time (second) on the above 6 images.

(144×176, 300), *WaterSurface* (128×160, 300), *ShoppingMall* (256×320, 100) and *ShopCorridor* (144×192, 200), where the numbers in the parentheses denote the frame size and the frame number. For each sequence with color frame size $h \times w$ and frame number k, we reshape it to a $(3hw) \times k$ matrix and use it in RPCA. To use SNN and TRPCA, we reshape the video to a $(hw) \times 3 \times k$ tensor[3]. The parameter of SNN in (4) is set to $\boldsymbol{\lambda} = [10,\ 0.1,\ 1] \times 20$ in this experiment.

3. We observe that this way of tensor construction performs well for TRPCA, despite one has some other ways.

Figure 8 shows the performance and running time comparison of RPCA, SNN and TRPCA on the four sequences. It can be seen that the low rank components identify the main illuminations as background, while the sparse parts correspond to the motion in the scene. Generally, our TRPCA performs the best. RPCA does not perform well on the *Hall* and *WaterSurface* sequences using the default parameter. Also, TRPCA is as efficient as RPCA and SNN requires much higher computational cost. The efficiency of TRPCA is benefited from our faster way for computing tensor SVT in Algorithm 3 which is the key step for solving TRPCA.



(a) Original (b) RPCA (c) SNN (d) TRPCA

	RPCA	SNN	TRPCA
Hall	**301.8**	1553.2	323.0
WaterSurface	250.1	887.3	**224.2**
ShoppingMall	**260.9**	744.0	372.4
ShopCorridor	**321.7**	1438.6	371.3

(e) Running time (seconds) comparison

Fig. 8: Background modeling results of four surveillance video sequences. (a) Original frames; (b)-(d) low rank and sparse components obtained by RPCA, SNN and TRPCA, respectively; (e) running time comparison.

6 Conclusions and Future Work

Based on the recently developed tensor-tensor product, which is a natural extension of the matrix-matrix product, we rigorously defined the tensor spectral norm, tensor nuclear norm and tensor average rank, such that their properties and relationships are consistent with the matrix cases. We then studied the Tensor Robust Principal Component (TRPCA) problem which aims to recover a low tubal rank tensor and a sparse tensor from their sum. We proved that under certain suitable assumptions, we can recover both the low-rank and the sparse components exactly by simply solving a convex program whose objective is a weighted combination of the tensor nuclear norm and the ℓ_1-norm. Benefitting from the "good" property of tensor nuclear norm, both our model and theoretical guarantee are natural extensions of RPCA. We also developed a more efficient method to compute the tensor singular value thresholding problem which is the key for solving TRPCA. Numerical experiments verify our theory and the results on images and videos demonstrate the effectiveness of our model.

There have some interesting future works. The work [7] generalizes the t-product using any invertible linear transform. With a proper choice of the invertible linear transform, it is possible to deduce a new tensor nuclear norm and solve the TRPCA problem. Beyond the convex models, the extensions to nonconvex cases are also important [21]. Finally, it is always interesting in using the developed tensor tools for real applications, e.g., image/video processing, web data analysis, and bioinformatics.

References

[1] K. Atkinson and W. Han. Theoretical numerical analysis: A functional analysis approach. *Texts in Applied Mathematics, Springer*, 2009.

[2] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optimization*, 2010.

[3] E. J. Candès, X. D. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3), 2011.

[4] Y. Chen. Incoherence-optimal matrix completion. *IEEE Trans. Information Theory*, 61(5):2909–2923, May 2015.

[5] Anandkumar, Anima and Deng, Yuan and Ge, Rong and Mobahi, Hossein. Homotopy analysis for tensor PCA. In *Conf. on Learning Theory*, 2017.

[6] Anandkumar, Anima and Jain, Prateek and Shi, Yang and Niranjan, Uma Naresh. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *Artificial Intelligence and Statistics*, pages 268–276, 2016.

[7] Kernfeld, Eric and Kilmer, Misha and Aeron, Shuchin. Tensor–tensor products with invertible linear transforms. *Linear Algebra and its Applications*, 485:545–570, 2015.

[8] Zhang, Anru and Xia, Dong. Tensor SVD: Statistical and Computational Limits. *IEEE Trans. Information Theory*, 2018.

[9] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[10] N. Hao, M. E. Kilmer, K. Braman, and R. C. Hoover. Facial recognition using tensor-tensor decompositions. *SIAM J. Imaging Sciences*, 6(1):437–463, 2013.

[11] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):45, 2013.

[12] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Springer, New York, 1993.

[13] B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable models for robust low-rank tensor completion. *Pacific J. Optimization*, 11(2):339–364, 2015.

[14] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM J. Matrix Analysis and Applications*, 34(1):148–172, 2013.

[15] M. E. Kilmer and C. D. Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011.

[16] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Recognition and Machine Intelligence*, 2013.

[18] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Recognition and Machine Intelligence*, 35(1):208–220, 2013.

[19] 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. Computer Vision and Pattern Recognition*. IEEE, 2016.
[20] C. Lu, J. Feng, S. Yan, and Z. Lin. A unified alternating direction method of multipliers by majorization minimization. *IEEE Trans. Pattern Recognition and Machine Intelligence*, 40(3):527–541, 2018.
[21] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin. Generalized singular value thresholding. In *Proc. AAAI Conf. Artificial Intelligence*, 2015.
[22] C. D. Martin, R. Shafer, and B. LaRue. An order-p tensor factorization with applications in imaging. *SIAM J. Scientific Computing*, 35(1):A474–A490, 2013.
[23] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE Int'l Conf. Computer Vision*, volume 2, pages 416–423. IEEE, 2001.
[24] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proc. Int'l Conf. Machine Learning*, pages 73–81, 2014.
[25] O. Rojo and H. Rojo. Some results on symmetric circulant matrices and on symmetric centrosymmetric matrices. *Linear algebra and its applications*, 392:211–233, 2004.
[26] B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems*, pages 2967–2975, 2013.
[27] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Processing*, 23(4):1678–1693, 2014.
[28] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.
[29] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
[30] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
[31] A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk. SpaRCS: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in Neural Information Processing Systems*, pages 1089–1097, 2011.
[32] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.
[33] Z. Zhang and S. Aeron. Exact tensor completion using t-SVD. *IEEE Trans. Signal Processing*, 65(6):1511–1526, 2017.
[34] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-SVD. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3842–3849. IEEE, 2014.

Appendix

At the following, we give the detailed proofs of Theorem 3.1, Theorem 3.2, and the main result in Theorem 4.1. Section A first gives some notations and properties which will be used in the proofs. Section B gives the proofs of Theorem 3.1 and 3.2 in our paper. Section C provides a way for the construction of the solution to the TRPCA model, and Section D proves that the constructed solution is optimal to the TRPCA problem. Section E gives the proofs of some lemmas which are used in Section D.

APPENDIX A
PRELIMINARIES

Beyond the notations introduced in the paper, we need some other notations used in the proofs. At the following, we define $\mathfrak{e}_{ijk} = \mathring{\mathbf{e}}_i * \dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^*$. Then we have $\boldsymbol{\mathcal{X}}_{ijk} = \langle \boldsymbol{\mathcal{X}}, \mathfrak{e}_{ijk} \rangle$. We define the projection

$$\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}}) = \sum_{ijk} \delta_{ijk} z_{ijk} \mathfrak{e}_{ijk},$$

where $\delta_{ijk} = 1_{(i,j,k)\in\boldsymbol{\Omega}}$, where $1_{(\cdot)}$ is the indicator function. Also $\boldsymbol{\Omega}^c$ denotes the complement of $\boldsymbol{\Omega}$ and $\mathcal{P}_{\boldsymbol{\Omega}^\perp}$ is the projection onto $\boldsymbol{\Omega}^c$. Denote $\boldsymbol{T}$ by the set

$$\boldsymbol{T} = \{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{Y}}^* + \boldsymbol{\mathcal{W}} * \boldsymbol{\mathcal{V}}^*,\ \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{W}} \in \mathbb{R}^{n\times r\times n_3}\}, \tag{45}$$

and by $\boldsymbol{T}^\perp$ its orthogonal complement. Then the projections onto $\boldsymbol{T}$ and $\boldsymbol{T}^\perp$ are respectively

$$\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{Z}} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^* - \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{Z}} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^*,$$

$$\begin{aligned}\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\mathcal{Z}}) =& \boldsymbol{\mathcal{Z}} - \mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\\ =& (\boldsymbol{\mathcal{I}}_{n_1} - \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^*) * \boldsymbol{\mathcal{Z}} * (\boldsymbol{\mathcal{I}}_{n_2} - \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^*),\end{aligned}$$

where $\boldsymbol{\mathcal{I}}_n$ denotes the $n \times n \times n_3$ identity tensor. Note that $\mathcal{P}_{\boldsymbol{T}}$ is self-adjoint. So we have

$$\begin{aligned}&\|\mathcal{P}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\\ =& \langle \mathcal{P}_{\boldsymbol{T}}(\mathfrak{e}_{ijk}), \mathfrak{e}_{ijk}\rangle\\ =& \langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \mathfrak{e}_{ijk} + \mathfrak{e}_{ijk} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^*, \mathfrak{e}_{ijk}\rangle\\ &- \langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \mathfrak{e}_{ijk} * \boldsymbol{\mathcal{V}} * \boldsymbol{\mathcal{V}}^*, \mathfrak{e}_{ijk}\rangle\end{aligned}$$

Note that

$$\begin{aligned}&\langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \mathfrak{e}_{ijk}, \mathfrak{e}_{ijk}\rangle\\ =& \langle \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i * \dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^*, \mathring{\mathbf{e}}_i * \dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^*\rangle\\ =& \langle \boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i, \boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i * (\dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^* * \mathring{\mathbf{e}}_j * \dot{\mathbf{e}}_k^*)\rangle\\ =& \langle \boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i, \boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i\rangle\\ =& \|\boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i\|_F^2,\end{aligned}$$

where we use the fact that $\dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^* * \mathring{\mathbf{e}}_j * \dot{\mathbf{e}}_k^* = \boldsymbol{\mathcal{I}}_1$, which is the $1 \times 1 \times n_3$ identity tensor. Therefore, it is easy to see that

$$\begin{aligned}&\|\mathcal{P}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\\ =& \|\boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i\|_F^2 + \|\boldsymbol{\mathcal{V}}^* * \mathring{\mathbf{e}}_j\|_F^2 - \|\boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i * \dot{\mathbf{e}}_k * \mathring{\mathbf{e}}_j^* * \boldsymbol{\mathcal{V}}\|_F^2,\\ \le& \|\boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i\|_F^2 + \|\boldsymbol{\mathcal{V}}^* * \mathring{\mathbf{e}}_j\|_F^2\\ \le& \frac{\mu r(n_1+n_2)}{n_1 n_2 n_3} \qquad (46)\\ =& \frac{2\mu r}{n n_3},\ \text{when } n_1 = n_2 = n. \qquad (47)\end{aligned}$$

where (46) uses the following tensor incoherence conditions

$$\max_{i=1,\cdots,n_1} \|\boldsymbol{\mathcal{U}}^* * \mathring{\mathbf{e}}_i\|_F \le \sqrt{\frac{\mu r}{n_1 n_3}}, \tag{48}$$

$$\max_{j=1,\cdots,n_2} \|\boldsymbol{\mathcal{V}}^* * \mathring{\mathbf{e}}_j\|_F \le \sqrt{\frac{\mu r}{n_2 n_3}}, \tag{49}$$

and

$$\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty \le \sqrt{\frac{\mu r}{n_1 n_2 n_3^2}}, \tag{50}$$

which are assumed to be satisfied in Theorem 4.1 in our manuscript.

APPENDIX B
PROOFS OF THEOREM 3.1 AND THEOREM 3.2

B.1 Proof of Theorem 3.1

Proof. To complete the proof, we need the conjugate function concept. The conjugate ϕ^* of a function $\phi : C \to \mathbb{R}$, where $C \subset \mathbb{R}^n$, is defined as

$$\phi^*(\boldsymbol{y}) = \sup\{\langle \boldsymbol{y}, \boldsymbol{x}\rangle - \phi(\boldsymbol{x}) | \boldsymbol{x} \in C\}.$$

Note that the conjugate of the conjugate, ϕ^{**}, is the convex envelope of the function ϕ. See Theorem 1.3.5 in [12], [9]. The proofs has two steps which compute ϕ^* and ϕ^{**}, respectively.

Step 1. *Computing* ϕ^*. For any $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, the conjugate function of the tensor average rank

$$\phi(\boldsymbol{\mathcal{A}}) = \text{rank}_\text{a}(\boldsymbol{\mathcal{A}}) = \frac{1}{n_3}\text{rank}(\texttt{bcirc}(\boldsymbol{\mathcal{A}})) = \frac{1}{n_3}\text{rank}(\bar{\boldsymbol{A}}),$$

on the set $S = \{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3} | \|\boldsymbol{\mathcal{A}}\| \le 1\}$ is

$$\begin{aligned}\phi^*(\boldsymbol{\mathcal{B}}) &= \sup_{\|\boldsymbol{\mathcal{A}}\|\le 1} \left(\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle - \text{rank}_\text{a}(\boldsymbol{\mathcal{A}})\right)\\ &= \sup_{\|\boldsymbol{\mathcal{A}}\|\le 1} \frac{1}{n_3}(\langle \bar{\boldsymbol{B}}, \bar{\boldsymbol{A}}\rangle - \text{rank}(\bar{\boldsymbol{A}})).\end{aligned}$$

Here $\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}} \in \mathbb{C}^{n_1 n_3\times n_2 n_3}$. Let $q = \min\{n_1 n_3, n_2 n_3\}$. By von Neumann's trace theorem,

$$\langle \bar{\boldsymbol{B}}, \bar{\boldsymbol{A}}\rangle \le \sum_{i=1}^{q} \sigma_i(\bar{\boldsymbol{B}})\sigma_i(\bar{\boldsymbol{A}}), \tag{51}$$

where $\sigma_i(\bar{\boldsymbol{A}})$ denotes the i-th largest singular value of $\bar{\boldsymbol{A}}$. Let $\bar{\boldsymbol{A}} = \bar{\boldsymbol{U}}_1\bar{\boldsymbol{S}}_1\bar{\boldsymbol{V}}_1^*$ and $\bar{\boldsymbol{B}} = \bar{\boldsymbol{U}}_2\bar{\boldsymbol{S}}_2\bar{\boldsymbol{V}}_2^*$ be the SVD of $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$, respectively. Note that the equality (51) holds when

$$\bar{\boldsymbol{U}}_1 = \bar{\boldsymbol{U}}_2 \text{ and } \bar{\boldsymbol{V}}_1 = \bar{\boldsymbol{V}}_2. \tag{52}$$

So we can pick $\bar{\boldsymbol{U}}_1$ and $\bar{\boldsymbol{V}}_1$ such that (52) holds to maximize $\langle \bar{\boldsymbol{B}}, \bar{\boldsymbol{A}}\rangle$. Note that the corresponding $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ of $\bar{\boldsymbol{U}}_1$ and $\bar{\boldsymbol{V}}_1$ respectively are real tensors and so is $\boldsymbol{\mathcal{A}}$ in this case. Thus, we have

$$\phi^*(\boldsymbol{\mathcal{B}}) = \sup_{\|\boldsymbol{\mathcal{A}}\|\le 1} \frac{1}{n_3}\left(\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{B}})\sigma_i(\bar{\boldsymbol{A}}) - \text{rank}(\bar{\boldsymbol{A}})\right).$$

If $\boldsymbol{\mathcal{A}} = 0$, then $\bar{\boldsymbol{A}} = 0$, and thus we have $\phi^*(\boldsymbol{\mathcal{B}}) = 0$ for all $\boldsymbol{\mathcal{B}}$. If $\text{rank}(\bar{\boldsymbol{A}}) = r$, $1 \le r \le q$, then $\phi^*(\boldsymbol{\mathcal{B}}) = \frac{1}{n_3}\left(\sum_{i=1}^{r}\sigma_i(\bar{\boldsymbol{B}}) - r\right)$. Hence $\phi^*(\boldsymbol{\mathcal{B}})$ can be expressed as

$$\begin{aligned}&n_3\cdot\phi^*(\boldsymbol{\mathcal{B}})\\ =&\max\left\{0, \sigma_1(\bar{\boldsymbol{B}}) - 1, \cdots, \sum_{i=1}^{r}\sigma_i(\bar{\boldsymbol{B}}) - r, \cdots, \sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{B}}) - q\right\}.\end{aligned}$$

The largest term in this set is the one that sums all positive $(\sigma_i(\bar{\boldsymbol{B}}) - 1)$ terms. Thus, we have

$$\begin{aligned}&\phi^*(\boldsymbol{\mathcal{B}})\\ =&\begin{cases}0, & \|\bar{\boldsymbol{B}}\| \le 1,\\ \frac{1}{n_3}\left(\sum_{i=1}^{r}\sigma_i(\bar{\boldsymbol{B}}) - r\right), & \sigma_r(\bar{\boldsymbol{B}}) > 1 \text{ and } \sigma_{r+1}(\bar{\boldsymbol{B}}) \le 1\end{cases}\\ =&\frac{1}{n_3}\sum_{i=1}^{q}(\sigma_i(\bar{\boldsymbol{B}}) - 1)_+.\end{aligned}$$

Note that above $\|\bar{\boldsymbol{B}}\| \le 1$ is equivalent to $\|\boldsymbol{\mathcal{B}}\| \le 1$.

Step 2. *Computing* ϕ^{**}. Now we compute the conjugate of ϕ^*, defined as

$$\begin{aligned}\phi^{**}(\boldsymbol{\mathcal{C}}) =& \sup_{\boldsymbol{\mathcal{B}}}(\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{B}}\rangle - \phi^*(\boldsymbol{\mathcal{B}}))\\ =& \sup_{\boldsymbol{\mathcal{B}}}\left(\frac{1}{n_3}\langle \bar{\boldsymbol{C}}, \bar{\boldsymbol{B}}\rangle - \phi^*(\boldsymbol{\mathcal{B}})\right),\end{aligned}$$

for all $\boldsymbol{\mathcal{C}} \in S$. As before, we can choose $\boldsymbol{\mathcal{B}}$ such that

$$\phi^{**}(\boldsymbol{\mathcal{C}}) = \sup_{\boldsymbol{\mathcal{B}}}\left(\frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\sigma_i(\bar{\boldsymbol{B}}) - \phi^*(\boldsymbol{\mathcal{B}})\right).$$

At the following, we consider two cases, $\|\boldsymbol{\mathcal{C}}\| > 1$ and $\|\boldsymbol{\mathcal{C}}\| \le 1$.

If $\|\boldsymbol{\mathcal{C}}\| > 1$, then $\sigma_1(\bar{\boldsymbol{C}}) = \|\bar{\boldsymbol{C}}\| = \|\boldsymbol{\mathcal{C}}\| > 1$. We can choose $\sigma_1(\bar{\boldsymbol{B}})$ large enough so that $\phi^{**}(\boldsymbol{\mathcal{C}}) \to \infty$. To see this, note that in

$$\phi^{**}(\boldsymbol{\mathcal{C}}) = \sup_{\boldsymbol{\mathcal{B}}}\frac{1}{n_3}\left(\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\sigma_i(\bar{\boldsymbol{B}}) - \left(\sum_{i=1}^{r}\sigma_i(\bar{\boldsymbol{B}}) - r\right)\right),$$

the coefficient of $\sigma_1(\bar{\boldsymbol{B}})$ is $\frac{1}{n_3}(\sigma_1(\bar{\boldsymbol{C}}) - 1)$ which is positive.

If $\|\boldsymbol{\mathcal{C}}\| \le 1$, then $\sigma_1(\bar{\boldsymbol{C}}) = \|\bar{\boldsymbol{C}}\| = \|\boldsymbol{\mathcal{C}}\| \le 1$. If $\|\boldsymbol{\mathcal{B}}\| = \|\bar{\boldsymbol{B}}\| \le 1$, then $\phi^*(\boldsymbol{\mathcal{B}}) = 0$ and the supremum is achieved for $\sigma_i(\bar{\boldsymbol{B}}) = 1$, $i = 1, \cdots, q$, yielding

$$\phi^{**}(\boldsymbol{\mathcal{C}}) = \frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}}) = \frac{1}{n_3}\|\bar{\boldsymbol{C}}\|_* = \|\boldsymbol{\mathcal{C}}\|_*.$$

If $\|\boldsymbol{\mathcal{C}}\| > 1$, we show that the argument of sup is is always smaller than $\|\boldsymbol{\mathcal{C}}\|_*$. By adding and subtracting the term $\frac{1}{n_3}\sum_i^q\sigma_i(\bar{\boldsymbol{C}})$ and rearranging the terms, we have

$$\begin{aligned}&\frac{1}{n_3}\left(\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\sigma_i(\bar{\boldsymbol{B}}) - \sum_{i=1}^{r}(\sigma_i(\bar{\boldsymbol{B}}) - 1)\right)\\ =&\frac{1}{n_3}\left(\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\sigma_i(\bar{\boldsymbol{B}}) - \sum_{i=1}^{r}(\sigma_i(\bar{\boldsymbol{B}}) - 1)\right)\\ &-\frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}}) + \frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\\ =&\frac{1}{n_3}\sum_{i=1}^{r}(\sigma_i(\bar{\boldsymbol{B}}) - 1)(\sigma_i(\bar{\boldsymbol{C}}) - 1)\\ &+\frac{1}{n_3}\sum_{i=r+1}^{q}(\sigma_i(\bar{\boldsymbol{B}}) - 1)\sigma_i(\bar{\boldsymbol{C}}) + \frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\\ <&\frac{1}{n_3}\sum_{i=1}^{q}\sigma_i(\bar{\boldsymbol{C}})\\ =&\|\boldsymbol{\mathcal{C}}\|_*.\end{aligned}$$

In a summary, we have shown that

$$\phi^{**}(\boldsymbol{\mathcal{C}}) = \|\boldsymbol{\mathcal{C}}\|_*,$$

over the set $S = \{\boldsymbol{\mathcal{C}} | \|\boldsymbol{\mathcal{C}}\| \le 1\}$. Thus, $\|\boldsymbol{\mathcal{C}}\|_*$ is the convex envelope of the tensor average rank $\text{rank}_\text{a}(\boldsymbol{\mathcal{C}})$ over S. □

B.2 Proof of Theorem 3.2

Proof. Let $\mathcal{G} \in \partial\|\mathcal{A}\|_*$. It is equivalent to the following statements [32]

$$\|\mathcal{A}\|_* = \langle \mathcal{G}, \mathcal{A} \rangle, \tag{53}$$
$$\|\mathcal{G}\| \leq 1. \tag{54}$$

So, to complete the proof, we only need to show that $\mathcal{G} = \mathcal{U} * \mathcal{V}^* + \mathcal{W}$, where $\mathcal{U}^* * \mathcal{W} = \mathbf{0}$, $\mathcal{W} * \mathcal{V} = \mathbf{0}$ and $\|\mathcal{W}\| \leq 1$, satisfies (53) and (54). First, we have

$$\begin{aligned}\langle \mathcal{G}, \mathcal{A} \rangle =& \langle \mathcal{U} * \mathcal{V}^* + \mathcal{W}, \mathcal{U} * \mathcal{S} * \mathcal{V}^* \rangle \\ =& \langle \mathcal{I}, \mathcal{S} \rangle + 0 \\ =& \|\mathcal{A}\|_*.\end{aligned}$$

Also, (54) is obvious when considering the property of $\mathcal{W}$. The proof is completed. □

APPENDIX C
DUAL CERTIFICATION

In this section, we first introduce conditions for $(\mathcal{L}_0, \mathcal{S}_0)$ to be the unique solution to TRPCA in subsection C.1. Then we construct a dual certificate in subsection C.2 which satisfies the conditions in subsection C.1, and thus our main result in Theorem 4.1 in our paper are proved.

C.1 Dual Certificates

Lemma C.1. *Assume that* $\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{P}_{\boldsymbol{T}}\| \leq \frac{1}{2}$ *and* $\lambda < \frac{1}{\sqrt{n_3}}$. *Then* $(\mathcal{L}_0, \mathcal{S}_0)$ *is the unique solution to the TRPCA problem if there is a pair* $(\mathcal{W}, \boldsymbol{F})$ *obeying*

$$(\mathcal{U} * \mathcal{V}^* + \mathcal{W}) = \lambda(\operatorname{sgn}(\mathcal{S}_0) + \boldsymbol{F} + \mathcal{P}_{\mathbf{\Omega}}\mathcal{D}),$$

with $\mathcal{P}_{\boldsymbol{T}}\mathcal{W} = \mathbf{0}$, $\|\mathcal{W}\| \leq \frac{1}{2}$, $\mathcal{P}_{\mathbf{\Omega}}\boldsymbol{F} = \mathbf{0}$ *and* $\|\boldsymbol{F}\|_\infty \leq \frac{1}{2}$, *and* $\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{D}\|_F \leq \frac{1}{4}$.

Proof. For any $\mathcal{H} \neq \mathbf{0}$, $(\mathcal{L}_0 + \mathcal{H}, \mathcal{S}_0 - \mathcal{H})$ is also a feasible solution. We show that its objective is larger than that at $(\mathcal{L}_0, \mathcal{S}_0)$, hence proving that $(\mathcal{L}_0, \mathcal{S}_0)$ is the unique solution. To do this, let $\mathcal{U} * \mathcal{V}^* + \mathcal{W}_0$ be an arbitrary subgradient of the tensor nuclear norm at $\mathcal{L}_0$, and $\operatorname{sgn}(\mathcal{S}_0) + \boldsymbol{F}_0$ be an arbitrary subgradient of the ℓ_1-norm at $\mathcal{S}_0$. Then we have

$$\begin{aligned}&\|\mathcal{L}_0 + \mathcal{H}\|_* + \lambda\|\mathcal{S}_0 - \mathcal{H}\|_1 \\ \geq& \|\mathcal{L}_0\|_* + \lambda\|\mathcal{S}_0\|_1 + \langle \mathcal{U} * \mathcal{V}^* + \mathcal{W}_0, \mathcal{H} \rangle \\ &- \lambda \langle \operatorname{sgn}(\mathcal{S}_0) + \boldsymbol{F}_0, \mathcal{H} \rangle.\end{aligned}$$

Now pick $\mathcal{W}_0$ such that $\langle \mathcal{W}_0, \mathcal{H} \rangle = \|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_*$ and $\langle \boldsymbol{F}_0, \mathcal{H} \rangle = -\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|$. We have

$$\begin{aligned}&\|\mathcal{L}_0 + \mathcal{H}\|_* + \lambda\|\mathcal{S}_0 - \mathcal{H}\|_1 \\ \geq& \|\mathcal{L}_0\|_* + \lambda\|\mathcal{S}_0\|_1 + \|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_* + \lambda\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_1 \\ &+ \langle \mathcal{U} * \mathcal{V}^* - \lambda \operatorname{sgn}(\mathcal{S}_0), \mathcal{H} \rangle.\end{aligned}$$

By assumption

$$\begin{aligned}&|\langle \mathcal{U} * \mathcal{V}^* - \lambda \operatorname{sgn}(\mathcal{S}_0), \mathcal{H} \rangle| \\ \leq& |\langle \mathcal{W}, \mathcal{H} \rangle| + \lambda |\langle \boldsymbol{F}, \mathcal{H} \rangle| + \lambda |\langle \mathcal{P}_{\mathbf{\Omega}}\mathcal{D}, \mathcal{H} \rangle| \\ \leq& \beta(\|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_* + \lambda\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_1) + \frac{\lambda}{4}\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{H}\|_F,\end{aligned}$$

where $\beta = \max(\|\mathcal{W}\|, \|\boldsymbol{F}\|_\infty) < \frac{1}{2}$. Thus

$$\begin{aligned}&\|\mathcal{L}_0 + \mathcal{H}\|_* + \lambda\|\mathcal{S}_0 - \mathcal{H}\|_1 \\ \geq& \|\mathcal{L}_0\|_* + \lambda\|\mathcal{S}_0\|_1 + \frac{1}{2}(\|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_* + \lambda\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_1) \\ &- \frac{\lambda}{4}\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{H}\|_F.\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{H}\|_F \leq& \|\mathcal{P}_{\mathbf{\Omega}}\mathcal{P}_{\boldsymbol{T}}\mathcal{H}\|_F + \|\mathcal{P}_{\mathbf{\Omega}}\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_F \\ \leq& \frac{1}{2}\|\mathcal{H}\|_F + \|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_F \\ \leq& \frac{1}{2}\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{H}\|_F + \frac{1}{2}\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_F + \|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_F.\end{aligned}$$

Thus

$$\begin{aligned}\|\mathcal{P}_{\mathbf{\Omega}}\mathcal{H}\|_F \leq& \|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_F + 2\|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_F \\ \leq& \|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_1 + 2\sqrt{n_3}\|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_*.\end{aligned}$$

In conclusion,

$$\begin{aligned}&\|\mathcal{L}_0 + \mathcal{H}\|_* + \lambda\|\mathcal{S}_0 - \mathcal{H}\|_1 \\ \geq& \|\mathcal{L}_0\|_* + \lambda\|\mathcal{S}_0\|_1 + \frac{1}{2}\left(1 - \lambda\sqrt{n_3}\right)\|\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{H}\|_* \\ &+ \frac{\lambda}{4}\|\mathcal{P}_{\mathbf{\Omega}^\perp}\mathcal{H}\|_1,\end{aligned}$$

where the last two terms are strictly positive when $\mathcal{H} \neq \mathbf{0}$. Thus, the proof is completed. □

Lemma C.1 implies that it is suffices to produce a dual certificate $\mathcal{W}$ obeying

$$\begin{cases}\mathcal{W} \in \boldsymbol{T}^\perp, \\ \|\mathcal{W}\| < \frac{1}{2}, \\ \|\mathcal{P}_{\mathbf{\Omega}}(\mathcal{U} * \mathcal{V}^* + \mathcal{W} - \lambda \operatorname{sgn}(\mathcal{S}_0))\|_F \leq \frac{\lambda}{4}, \\ \|\mathcal{P}_{\mathbf{\Omega}^\perp}(\mathcal{U} * \mathcal{V}^* + \mathcal{W})\|_\infty < \frac{\lambda}{2}.\end{cases} \tag{55}$$

C.2 Dual Certification via the Golfing Scheme

In this subsection, we show how to construct a dual certificate obeying (55). Before we introduce our construction, our model assumes that $\mathbf{\Omega} \sim \text{Ber}(\rho)$, or equivalently that $\mathbf{\Omega}^c \sim \text{Ber}(1-\rho)$. Now the distribution of $\mathbf{\Omega}^c$ is the same as that of $\mathbf{\Omega}^c = \mathbf{\Omega}_1 \cup \mathbf{\Omega}_2 \cup \cdots \cup \mathbf{\Omega}_{j_0}$, where each $\mathbf{\Omega}_j$ follows the Bernoulli model with parameter q, which satisfies

$$\mathbb{P}((i,j,k) \in \mathbf{\Omega}) = \mathbb{P}(\text{Bin}(j_0, q) = 0) = (1-q)^{j_0},$$

so that the two models are the same if $\rho = (1-q)^{j_0}$. Note that because of overlaps between the $\mathbf{\Omega}_j$'s, $q \geq (1-\rho)/j_0$.

Now, we construct a dual certificate

$$\mathcal{W} = \mathcal{W}^{\mathcal{L}} + \mathcal{W}^{\mathcal{S}}, \tag{56}$$

where each component is as follows:

1) Construction of $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}$ via the Golfing Scheme. Let $j_0 = 2\log(nn_3)$ and $\boldsymbol{\Omega}_j$, $j = 1, \cdots, j_0$, be defined as previously described so that $\boldsymbol{\Omega}^c = \cup_{1\le j\le j_0}\boldsymbol{\Omega}_j$. Then define

$$\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}} = \boldsymbol{\mathcal{P}}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{Y}}_{j_0}, \tag{57}$$

where

$$\boldsymbol{\mathcal{Y}}_j = \boldsymbol{\mathcal{Y}}_{j-1} + q^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* - \boldsymbol{\mathcal{Y}}_{j-1}),\ \boldsymbol{\mathcal{Y}}_0 = \boldsymbol{0}.$$

2) Construction of $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}$ via the Method of Least Squares. Assume that $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| < 1/2$. Then, $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\| < 1/4$, and thus, the operator $\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}$ mapping $\boldsymbol{\Omega}$ onto itself is invertible; we denote its inverse by $(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}$. We then set

$$\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}} = \lambda\boldsymbol{\mathcal{P}}_{\boldsymbol{T}^\perp}(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}\operatorname{sgn}(\boldsymbol{\mathcal{S}}_0). \tag{58}$$

This is equivalent to

$$\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}} = \lambda\boldsymbol{\mathcal{P}}_{\boldsymbol{T}^\perp}\sum_{k\ge 0}(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^k\operatorname{sgn}(\boldsymbol{\mathcal{S}}_0).$$

Since both $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}$ and $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}$ belong to $\boldsymbol{T}^\perp$ and $\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}} = \lambda\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{P}}_{\boldsymbol{T}})(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}-\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}\operatorname{sgn}(\boldsymbol{\mathcal{S}}_0) = \lambda\operatorname{sgn}(\boldsymbol{\mathcal{S}}_0)$, we will establish that $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}} + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}$ is a valid dual certificate if it obeys

$$\begin{cases}\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}} + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}\| < \frac{1}{2},\\ \|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}})\|_F \le \frac{\lambda}{4},\\ \|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}} + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}})\|_\infty < \frac{\lambda}{2}.\end{cases} \tag{59}$$

This can be achieved by using the following two key lemmas:

Lemma C.2. *Assume that $\boldsymbol{\Omega} \sim Ber(\rho)$ with parameter $\rho \le \rho_s$ for some $\rho_s > 0$. Set $j_0 = 2\lceil\log(nn_3)\rceil$ (use $\log(n_{(1)}n_3)$ for the tensors of rectangular frontal slice). Then, the tensor $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}$ obeys*

(a) $\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}\| < \frac{1}{4}$,
(b) $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}})\|_F < \frac{\lambda}{4}$,
(c) $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}})\|_\infty < \frac{\lambda}{4}$.

Lemma C.3. *Assume that $\boldsymbol{\mathcal{S}}_0$ is supported on a set $\boldsymbol{\Omega}$ sampled as in Lemma C.2, and that the signs of $\boldsymbol{\mathcal{S}}_0$ are independent and identically distributed symmetric (and independent of $\boldsymbol{\Omega}$). Then, the tensor $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}$ (58) obeys*

(a) $\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}\| < \frac{1}{4}$,
(b) $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}\|_\infty < \frac{\lambda}{4}$.

So the left task is to prove Lemma C.2 and Lemma C.3, which are given in Section D.

APPENDIX D
PROOFS OF DUAL CERTIFICATION

This section gives the proofs of Lemma C.2 and Lemma C.3. To do this, we first introduce some lemmas with their proofs given in Section E.

Lemma D.1. *For the Bernoulli sign variable $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n\times n\times n_3}$ defined as*

$$\boldsymbol{\mathcal{M}}_{ijk} = \begin{cases}1, & w.p.\ \rho/2,\\ 0, & w.p.\ 1-\rho,\\ -1, & w.p.\ \rho/2,\end{cases} \tag{60}$$

where $\rho > 0$, there exists a function $\varphi(\rho)$ satisfying $\lim_{\rho\to 0^+}\varphi(\rho) = 0$, such that the following statement holds with with large probability,

$$\|\boldsymbol{\mathcal{M}}\| \le \varphi(\rho)\sqrt{nn_3}.$$

Lemma D.2. *Suppose $\boldsymbol{\Omega} \sim Ber(\rho)$. Then with high probability,*

$$\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}} - \rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \epsilon,$$

provided that $\rho \ge C_0\epsilon^{-2}(\mu r\log(nn_3))/(nn_3)$ for some numerical constant $C_0 > 0$. For the tensor of rectangular frontal slice, we need $\rho \ge C_0\epsilon^{-2}(\mu r\log(n_{(1)}n_3))/(n_{(2)}n_3)$.

Corollary D.3. *Assume that $\boldsymbol{\Omega} \sim Ber(\rho)$, then $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\|^2 \le \rho + \epsilon$, provided that $1-\rho \ge C\epsilon^{-2}(\mu r\log(nn_3))/(nn_3)$, where C is as in Lemma D.2. For the tensor with frontal slice, the modification is as in Lemma D.2.*

Note that this corollary shows that $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le 1/2$, provided $|\boldsymbol{\Omega}|$ is not too large.

Lemma D.4. *Suppose that $\boldsymbol{\mathcal{Z}} \in \boldsymbol{T}$ is a fixed tensor, and $\boldsymbol{\Omega} \sim Ber(\rho)$. Then, with high probability,*

$$\|\boldsymbol{\mathcal{Z}} - \rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{Z}}\|_\infty \le \epsilon\|\boldsymbol{\mathcal{Z}}\|_\infty, \tag{61}$$

provided that $\rho \ge C_0\epsilon^{-2}(\mu r\log(nn_3))/(nn_3)$ (for the tensor of rectangular frontal slice, $\rho \ge C_0\epsilon^{-2}(\mu r\log(n_{(1)}n_3))/(n_{(2)}n_3)$) for some numerical constant $C_0 > 0$.

Lemma D.5. *Suppose $\boldsymbol{\mathcal{Z}}$ is fixed, and $\boldsymbol{\Omega} \sim Ber(\rho)$. Then, with high probability,*

$$\|(\boldsymbol{\mathcal{I}} - \rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})\boldsymbol{\mathcal{Z}}\| \le \sqrt{\frac{C_0nn_3\log(nn_3)}{\rho}}\|\boldsymbol{\mathcal{Z}}\|_\infty, \tag{62}$$

for some numerical constant $C_0 > 0$ provided that $\rho \ge C_0\log(nn_3)/(nn_3)$ (or $\rho \ge C_0\log(n_{(1)}n_3)/(n_{(2)}n_3)$ for the tensors with rectangular frontal slice).

D.1 Proof of Lemma C.2

Proof. We first introduce some notations. Set $\boldsymbol{\mathcal{Z}}_j = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* - \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{Y}}_j$ obeying

$$\boldsymbol{\mathcal{Z}}_j = (\boldsymbol{\mathcal{P}}_{\boldsymbol{T}} - q^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}})\boldsymbol{\mathcal{Z}}_{j-1}.$$

So $\boldsymbol{\mathcal{Z}}_j \in \boldsymbol{\mathcal{T}}$ for all $j \ge 0$. Also, note that when

$$q \ge C_0\epsilon^{-2}\frac{\mu r\log(nn_3)}{nn_3}, \tag{63}$$

or for the tensors with rectangular frontal slices $q \ge C_0\epsilon^{-2}\frac{\mu r\log(n_{(1)}n_3)}{n_{(2)}n_3}$, we have

$$\|\boldsymbol{\mathcal{Z}}_j\|_\infty \le \epsilon\|\boldsymbol{\mathcal{Z}}_{j-1}\|_\infty \le \epsilon^j\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty, \tag{64}$$

by Lemma D.4 and

$$\|\boldsymbol{\mathcal{Z}}_j\|_F \leq \epsilon\|\boldsymbol{\mathcal{Z}}_{j-1}\|_F \leq \epsilon^j\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_F \leq \epsilon^j\sqrt{r}. \tag{65}$$

We assume $\epsilon \leq e^{-1}$.

1. Proof of (a). Note that $\boldsymbol{\mathcal{Y}}_{j_0} = \sum_j q^{-1}\mathcal{P}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{Z}}_{j-1}$. We have

$$\begin{aligned}
\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}\| =&\|\mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{Y}}_{j_0}\| \leq \sum_j \|q^{-1}\mathcal{P}_{\boldsymbol{T}^\perp}\mathcal{P}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{Z}}_{j-1}\| \\
\leq& \sum_j \|\mathcal{P}_{\boldsymbol{T}^\perp}(q^{-1}\mathcal{P}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{Z}}_{j-1} - \boldsymbol{\mathcal{Z}}_{j-1})\| \\
\leq& \sum_j \|q^{-1}\mathcal{P}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{Z}}_{j-1} - \boldsymbol{\mathcal{Z}}_{j-1}\| \\
\leq& C_0'\sqrt{\frac{nn_3\log(nn_3)}{q}}\sum_j \|\boldsymbol{\mathcal{Z}}_{j-1}\|_\infty \\
\leq& C_0'\sqrt{\frac{nn_3\log(nn_3)}{q}}\sum_j \epsilon^{j-1}\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty \\
\leq& C_0'(1-\epsilon)^{-1}\sqrt{\frac{nn_3\log(nn_3)}{q}}\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty.
\end{aligned}$$

The fourth step is from Lemma D.5 and the fifth is from (64). Now by using (63) and (50), we have

$$\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}}\| \leq C'\epsilon,$$

for some numerical constant C'.

2. Proof of (b). Since $\mathcal{P}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{Y}}_{j_0} = \mathbf{0}$, $\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{Y}}_{j_0}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* - \mathcal{P}_{\boldsymbol{T}}\boldsymbol{\mathcal{Y}}_{j_0}) = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}}_{j_0})$, and it follows from (65) that

$$\|\boldsymbol{\mathcal{Z}}_{j_0}\|_F \leq \epsilon^{j_0}\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_F \leq \epsilon^{j_0}\sqrt{r}.$$

Since $\epsilon \leq e^{-1}$ and $j_0 \geq 2\log(nn_3)$, $\epsilon^{j_0} \leq (nn_3)^{-2}$ and this proves the claim.

3. Proof of (c). We have $\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{L}}} = \boldsymbol{\mathcal{Z}}_{j_0} + \boldsymbol{\mathcal{Y}}_{j_0}$ and know that $\boldsymbol{\mathcal{Y}}_{j_0}$ is supported on $\boldsymbol{\Omega}^c$. Therefore, since $\|\boldsymbol{\mathcal{Z}}_{j_0}\|_F \leq \lambda/8$. We only need to show that $\|\boldsymbol{\mathcal{Y}}_{j_0}\|_\infty \leq \lambda/8$. Indeed,

$$\begin{aligned}
\|\boldsymbol{\mathcal{Y}}_{j_0}\|_\infty \leq& q^{-1}\sum_j \|\mathcal{P}_{\boldsymbol{\Omega}_j}\boldsymbol{\mathcal{Z}}_{j-1}\|_\infty \\
\leq& q^{-1}\sum_j \|\boldsymbol{\mathcal{Z}}_{j-1}\|_\infty \\
\leq& q^{-1}\sum_j \epsilon^{j-1}\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty.
\end{aligned}$$

Since $\|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^*\|_\infty \leq \sqrt{\frac{\mu r}{n^2n_3^2}}$, this gives

$$\|\boldsymbol{\mathcal{Y}}_{j_0}\|_\infty \leq C'\frac{\epsilon^2}{\sqrt{\mu r(\log(nn_3))^2}},$$

for some numerical constant C' whenever q obeys (63). Since $\lambda = 1/\sqrt{nn_3}$, $\|\boldsymbol{\mathcal{Y}}_{j_0}\|_\infty \leq \lambda/8$ if

$$\epsilon \leq C\left(\frac{\mu r(\log(nn_3))^2}{nn_3}\right)^{1/4}.$$

The claim is proved by using (63), (50) and sufficiently small ϵ (provided that ρ_r is sufficiently small. Note that everything is consistent since $C_0\epsilon^{-2}\frac{\mu r\log(nn_3)}{nn_3} < 1$. □

D.2 Proof of Lemma C.3

Proof. We denote $\boldsymbol{\mathcal{M}} = \operatorname{sgn}(\boldsymbol{\mathcal{S}}_0)$ distributed as

$$\boldsymbol{\mathcal{M}}_{ijk} = \begin{cases} 1, & \text{w.p. } \rho/2, \\ 0, & \text{w.p. } 1-\rho, \\ -1, & \text{w.p. } \rho/2. \end{cases}$$

Note that for any $\sigma > 0$, $\{\|\mathcal{P}_{\boldsymbol{\Omega}}\mathcal{P}_{\boldsymbol{T}}\| \leq \sigma\}$ holds with high probability provided that ρ is sufficiently small, see Corollary D.3.

1. Proof of (a). By construction,

$$\begin{aligned}
\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}} =&\lambda\mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{M}} + \lambda\mathcal{P}_{\boldsymbol{T}^\perp}\sum_{k\geq 1}(\mathcal{P}_{\boldsymbol{\Omega}}\mathcal{P}_{\boldsymbol{T}}\mathcal{P}_{\boldsymbol{\Omega}})^k\boldsymbol{\mathcal{M}} \\
:=&\mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{W}}_0^{\boldsymbol{\mathcal{S}}} + \mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{W}}_1^{\boldsymbol{\mathcal{S}}}.
\end{aligned}$$

Note that $\|\mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{W}}_0^{\boldsymbol{\mathcal{S}}}\| \leq \|\boldsymbol{\mathcal{W}}_0^{\boldsymbol{\mathcal{S}}}\| = \lambda\|\boldsymbol{\mathcal{M}}\|$ and $\|\mathcal{P}_{\boldsymbol{T}^\perp}\boldsymbol{\mathcal{W}}_1^{\boldsymbol{\mathcal{S}}}\| \leq \|\boldsymbol{\mathcal{W}}_1^{\boldsymbol{\mathcal{S}}}\| = \lambda\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\|$, where $\boldsymbol{\mathcal{R}} = \sum_{k\geq 1}(\mathcal{P}_{\boldsymbol{\Omega}}\mathcal{P}_{\boldsymbol{T}}\mathcal{P}_{\boldsymbol{\Omega}})^k$. Now, we will respectively show that $\lambda\|\boldsymbol{\mathcal{M}}\|$ and $\lambda\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\|$ are small enough when ρ is sufficiently small for $\lambda = 1/\sqrt{nn_3}$. Therefor, $\|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}\| \leq 1/4$.

1) Bound $\lambda\|\boldsymbol{\mathcal{M}}\|$.

By using Lemma D.1 directly, we have that $\lambda\|\boldsymbol{\mathcal{M}}\| \leq \varphi(\rho)$ is sufficiently small given $\lambda = 1/\sqrt{nn_3}$ and ρ is sufficiently small.

2) Bound $\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\|$.

For simplicity, let $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})$. We have

$$\|\boldsymbol{\mathcal{Z}}\| = \|\bar{\boldsymbol{Z}}\| = \sup_{\boldsymbol{x}\in\mathbb{S}^{nn_3-1}}\|\bar{\boldsymbol{Z}}\boldsymbol{x}\|_2. \tag{66}$$

The optimal $\boldsymbol{x}$ to (66) is an eigenvector of $\bar{\boldsymbol{Z}}^*\bar{\boldsymbol{Z}}$. Since $\bar{\boldsymbol{Z}}$ is a block diagonal matrix, the optimal $\boldsymbol{x}$ has a block sparse structure, i.e., $\boldsymbol{x} \in B = \{\boldsymbol{x} \in \mathbb{R}^{nn_3} | \boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_i^\top \cdots, \boldsymbol{x}_{n_3}^\top]$, with $\boldsymbol{x}_i \in \mathbb{R}^n$, and there exists j such that $\boldsymbol{x}_j \neq \mathbf{0}$ and $\boldsymbol{x}_i = \mathbf{0}, i \neq j\}$. Note that $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}_j\|_2 = 1$. Let N be the $1/2$-net for $\mathbb{S}^{n-1}$ of size at most 5^n (see Lemma 5.2 in [30]). Then the $1/2$-net, denoted as N', for B has the size at most $n_3 \cdot 5^n$. We have

$$\begin{aligned}
\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\| =&\|\texttt{bdiag}(\overline{\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})})\| \\
=& \sup_{\boldsymbol{x},\boldsymbol{y}\in B}\left\langle \boldsymbol{x}, \texttt{bdiag}(\overline{\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})})\boldsymbol{y}\right\rangle \\
=& \sup_{\boldsymbol{x},\boldsymbol{y}\in B}\left\langle \boldsymbol{x}\boldsymbol{y}^*, \texttt{bdiag}(\overline{\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})})\right\rangle \\
=& \sup_{\boldsymbol{x},\boldsymbol{y}\in B}\left\langle \texttt{bdiag}^*(\boldsymbol{x}\boldsymbol{y}^*), \overline{\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})}\right\rangle,
\end{aligned}$$

where $\texttt{bdiag}^*$, the joint operator of $\texttt{bdiag}$, maps the block diagonal matrix $\boldsymbol{x}\boldsymbol{y}^*$ to a tensor of size $n \times n \times n_3$. Let $\boldsymbol{\mathcal{Z}}' = \texttt{bdiag}^*(\boldsymbol{x}\boldsymbol{y}^*)$ and $\boldsymbol{\mathcal{Z}} = \texttt{ifft}(\boldsymbol{\mathcal{Z}}', [\,], 3)$. We have

$$\begin{aligned}
\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\| =& \sup_{\boldsymbol{x},\boldsymbol{y}\in B}\left\langle \boldsymbol{\mathcal{Z}}', \overline{\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})}\right\rangle \\
=& \sup_{\boldsymbol{x},\boldsymbol{y}\in B} n_3\left\langle \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\right\rangle \\
=& \sup_{\boldsymbol{x},\boldsymbol{y}\in B} n_3\left\langle \boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{Z}}), \boldsymbol{\mathcal{M}}\right\rangle \\
\leq& \sup_{\boldsymbol{x},\boldsymbol{y}\in N'} 4n_3\left\langle \boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{Z}}), \boldsymbol{\mathcal{M}}\right\rangle.
\end{aligned}$$

For a fixed pair $(\boldsymbol{x},\boldsymbol{y})$ of unit-normed vectors, define the random variable

$$X(\boldsymbol{x},\boldsymbol{y}) = \langle 4n_3\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{Z}}), \boldsymbol{\mathcal{M}}\rangle .$$

Conditional on $\boldsymbol{\Omega} = \text{supp}(\boldsymbol{\mathcal{M}})$, the signs of $\boldsymbol{\mathcal{M}}$ are independent and identically distributed symmetric and Hoeffding's inequality gives

$$\mathbb{P}(|X(\boldsymbol{x},\boldsymbol{y})| > t|\boldsymbol{\Omega}) \le 2\exp\left(\frac{-2t^2}{\|4n_3\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{Z}})\|_F^2}\right).$$

Note that $\|4n_3\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{Z}}))\|_F \le 4n_3\|\boldsymbol{\mathcal{R}}\|\|\boldsymbol{\mathcal{Z}}\|_F = 4\sqrt{n_3}\|\boldsymbol{\mathcal{R}}\|\|\boldsymbol{\mathcal{Z}}'\|_F = 4\sqrt{n_3}\|\boldsymbol{\mathcal{R}}\|$. Therefore, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{x},\boldsymbol{y}\in N'}|X(\boldsymbol{x},\boldsymbol{y})| > t|\boldsymbol{\Omega}\right) \le 2|N'|^2\exp\left(-\frac{t^2}{8n_3\|\boldsymbol{\mathcal{R}}\|^2}\right).$$

Hence,

$$\mathbb{P}\left(\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\| > t|\boldsymbol{\Omega}\right) \le 2|N'|^2\exp\left(-\frac{t^2}{8n_3\|\boldsymbol{\mathcal{R}}\|^2}\right).$$

On the event $\{\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \sigma\}$,

$$\|\boldsymbol{\mathcal{R}}\| \le \sum_{k\ge 1}\sigma^{2k} = \frac{\sigma^2}{1-\sigma^2},$$

and, therefore, unconditionally,

$$\begin{aligned}&\mathbb{P}\left(\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\| > t\right)\\ \le& 2|N'|^2\exp\left(-\frac{\gamma^2t^2}{8n_3}\right) + \mathbb{P}\left(\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \ge \sigma\right),\ \gamma = \frac{1-\sigma^2}{2\sigma^2}\\ =& 2n_3^2\cdot 5^{2n}\exp\left(-\frac{\gamma^2t^2}{8n_3}\right) + \mathbb{P}\left(\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \ge \sigma\right).\end{aligned}$$

Let $t = c\sqrt{nn_3}$, where c can be a small absolute constant. Then the above inequality implies that $\|\boldsymbol{\mathcal{R}}(\boldsymbol{\mathcal{M}})\| \le t$ with high probability.

2. Proof of (b) Observe that

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}} = -\lambda\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}\boldsymbol{\mathcal{M}}.$$

Now for $(i,j,k) \in \boldsymbol{\Omega}^c$, $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}_{ijk} = \left\langle \boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}, \mathfrak{e}_{ijk}\right\rangle$, and we have $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}_{ijk} = \lambda\langle \boldsymbol{\mathcal{Q}}(i,j,k), \boldsymbol{\mathcal{M}}\rangle$, where $\boldsymbol{\mathcal{Q}}(i,j,k)$ is the tensor $-(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})$. Conditional on $\boldsymbol{\Omega} = \text{supp}(\boldsymbol{\mathcal{M}})$, the signs of $\boldsymbol{\mathcal{M}}$ are independent and identically distributed symmetric, and the Hoeffding's inequality gives

$$\mathbb{P}(|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}_{ijk}| > t\lambda|\boldsymbol{\Omega}) \le 2\exp\left(-\frac{2t^2}{\|\boldsymbol{\mathcal{Q}}(i,j,k)\|_F^2}\right),$$

and

$$\begin{aligned}&\mathbb{P}(\sup_{i,j,k}|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}_{ijk}| > t\lambda/n_3|\boldsymbol{\Omega})\\ \le& 2n^2n_3\exp\left(-\frac{2t^2}{\sup_{i,j,k}\|\boldsymbol{\mathcal{Q}}(i,j,k)\|_F^2}\right).\end{aligned}$$

By using (47), we have

$$\begin{aligned}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F \le& \|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\|\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F\\ \le& \sigma\sqrt{\frac{2\mu r}{nn_3}},\end{aligned}$$

on the event $\{\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \sigma\}$. On the same event, we have $\|(\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})^{-1}\| \le (1-\sigma^2)^{-1}$ and thus $\|\boldsymbol{\mathcal{Q}}(i,j,k)\|_F^2 \le \frac{2\sigma^2}{(1-\sigma^2)^2}\frac{\mu r}{nn_3}$. Then, unconditionally,

$$\begin{aligned}&\mathbb{P}\left(\sup_{i,j,k}|\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{S}}}_{ijk}| > t\lambda\right)\\ \le& 2n^2n_3\exp\left(-\frac{nn_3\gamma^2t^2}{\mu r}\right) + \mathbb{P}(\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \ge \sigma),\end{aligned}$$

where $\gamma = \frac{(1-\sigma^2)^2}{2\sigma^2}$. This proves the claim when $\mu r < \rho_r' nn_3\log(nn_3)^{-1}$ and ρ_r' is sufficiently small. □

APPENDIX E
PROOFS OF SOME LEMMAS

Lemma E.1. *[29] Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random $n_1\times n_2$ matrices that satisfy the assumption $\mathbb{E}\boldsymbol{Z}_k = \boldsymbol{0}$ and $\|\boldsymbol{Z}_k\| \le R$ almost surely. Let $\sigma^2 = \max\{\|\sum_k\mathbb{E}[\boldsymbol{Z}_k\boldsymbol{Z}_k^*]\|, \max\{\|\sum_k\mathbb{E}[\boldsymbol{Z}_k^*\boldsymbol{Z}_k]\|\}$. Then, for any $t\ge 0$, we have*

$$\begin{aligned}\mathbb{P}\left[\left\|\sum_k\boldsymbol{Z}_k\right\| \ge t\right] \le& (n_1+n_2)\exp\left(-\frac{t^2}{2\sigma^2+\frac{2}{3}Rt}\right)\\ \le& (n_1+n_2)\exp\left(-\frac{3t^2}{8\sigma^2}\right),\ \textit{for } t\le\frac{\sigma^2}{R}.\end{aligned}$$

Or, for any $c>0$, we have

$$\left\|\sum_k\boldsymbol{Z}_k\right\| \ge 2\sqrt{c\sigma^2\log(n_1+n_2)} + cB\log(n_1+n_2),$$

with probability at least $1-(n_1+n_2)^{1-c}$.

E.1 Proof of Lemma D.1

Proof. The proof has three steps.

Step 1: Approximation. We first introduce some notations. Let $\boldsymbol{f}_i^*$ be the i-th row of $\boldsymbol{F}_{n_3}$, and $\boldsymbol{M}^H = \begin{bmatrix}\boldsymbol{M}_1^H\\ \boldsymbol{M}_2^H\\ \vdots\\ \boldsymbol{M}_n^H\end{bmatrix} \in \mathbb{R}^{nn_3\times n}$ be a matrix unfolded by $\boldsymbol{\mathcal{M}}$, where $\boldsymbol{M}_i^H \in \mathbb{R}^{n_3\times n}$ is the i-th horizontal slice of $\boldsymbol{\mathcal{M}}$, i.e., $[\boldsymbol{M}_i^H]_{kj} = \boldsymbol{\mathcal{M}}_{ikj}$. Consider that $\bar{\boldsymbol{\mathcal{M}}} = \texttt{fft}(\boldsymbol{\mathcal{M}}, [\,], 3)$, we have

$$\bar{\boldsymbol{M}}_i = \begin{bmatrix}\boldsymbol{f}_i^*\boldsymbol{M}_1^H\\ \boldsymbol{f}_i^*\boldsymbol{M}_2^H\\ \vdots\\ \boldsymbol{f}_i^*\boldsymbol{M}_n^H\end{bmatrix},$$

where $\bar{\boldsymbol{M}}_i \in \mathbb{R}^{n\times n}$ is the i-th frontal slice of $\boldsymbol{\mathcal{M}}$. Note that

$$\|\boldsymbol{\mathcal{M}}\| = \|\bar{\boldsymbol{M}}\| = \max_{i=1,\cdots,n_3}\|\bar{\boldsymbol{M}}_i\|. \tag{67}$$

Let N be the $1/2$-net for $\mathbb{S}^{n-1}$ of size at most 5^n (see Lemma 5.2 in [30]). Then Lemma 5.3 in [30] gives

$$\|\bar{\boldsymbol{M}}_i\| \le 2\max_{\boldsymbol{x}\in N}\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2. \tag{68}$$

So we consider to bound $\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2$.

Step 2: Concentration. We can express $\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2$ as a sum of independent random variables

$$\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2=\sum_{j=1}^{n}(\boldsymbol{f}_i^*\boldsymbol{M}_j^H\boldsymbol{x})^2:=\sum_{j=1}^{n}z_j^2, \tag{69}$$

where $z_j=\langle\boldsymbol{M}_j^H,\boldsymbol{f}_i\boldsymbol{x}^*\rangle$, $j=1,\cdots,n$, are independent sub-gaussian random variables with $\mathbb{E}z_j^2=\rho\|\boldsymbol{f}_i\boldsymbol{x}^*\|_F^2=\rho n_3$. Using (60), we have

$$|[\boldsymbol{M}_j^H]_{kl}|=\begin{cases}1, & \text{w.p. } \rho,\\ 0, & \text{w.p. } 1-\rho.\end{cases}$$

Thus, the sub-gaussian norm of $[\boldsymbol{M}_j^H]_{kl}$, denoted as $\|\cdot\|_{\psi_2}$, is

$$\begin{aligned}\|[\boldsymbol{M}_j^H]_{kl}\|_{\psi_2}=&\sup_{p\geq 1}p^{-\frac{1}{2}}(\mathbb{E}[|[\boldsymbol{M}_j^H]_{kl}|^p])^{\frac{1}{p}}\\ =&\sup_{p\geq 1}p^{-\frac{1}{2}}\rho^{\frac{1}{p}}.\end{aligned}$$

Define the function $\phi(x)=x^{-\frac{1}{2}}\rho^{\frac{1}{x}}$ on $[1,+\infty)$. The only stationary point occurs at $x^*=\log\rho^{-2}$. Thus,

$$\begin{aligned}\phi(x)\leq&\max(\phi(1),\phi(x^*))\\ =&\max\left(\rho,(\log\rho^{-2})^{-\frac{1}{2}}\rho^{\frac{1}{\log\rho^{-2}}}\right)\\ :=&\psi(\rho).\end{aligned}\tag{70}$$

Therefore, $\|[\boldsymbol{M}_j^H]_{kl}\|_{\psi_2}\leq\psi(\rho)$. Consider that z_j is a sum of independent centered sub-gaussian random variables $[\boldsymbol{M}_j^H]_{kl}$'s, by using Lemma 5.9 in [30], we have $\|z_j\|_{\psi_2}^2\leq c_1(\psi(\rho))^2n_3$, where c_1 is an absolute constant. Therefore, by Remark 5.18 and Lemma 5.14 in [30], $z_j^2-\rho n_3$ are independent centered sub-exponential random variables with $\|z_j^2-\rho n_3\|_{\psi_1}\leq 2\|z_j\|_{\psi_1}^2\leq 4\|z_j\|_{\psi_2}^2\leq 4c_1(\psi(\rho))^2n_3$.

Now, we use an exponential deviation inequality, Corollary 5.17 in [30], to control the sum (69). We have

$$\begin{aligned}&\mathbb{P}\left(|\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2-\rho nn_3|\geq tn\right)\\ =&\mathbb{P}\left(\left|\sum_{j=1}^{n}(z_j^2-\rho n_3)\right|\geq tn\right)\\ \leq&2\exp\left(-c_2n\min\left(\left(\frac{t}{4c_1(\psi(\rho))^2n_3}\right)^2,\frac{t}{4c_1(\psi(\rho))^2n_3}\right)\right),\end{aligned}$$

where $c_2>0$. Let $t=c_3(\psi(\rho))^2n_3$ for some absolute constant c_3, we have

$$\begin{aligned}&\mathbb{P}\left(|\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2-\rho nn_3|\geq c_3(\psi(\rho))^2nn_3\right)\\ \leq&2\exp\left(-c_2n\min\left(\left(\frac{c_3}{4c_1}\right)^2,\frac{c_3}{4c_1}\right)\right).\end{aligned}$$

Step 3 Union bound. Taking the union bound over all $\boldsymbol{x}$ in the net N of cardinality $|N|\leq 5^n$, we obtain

$$\begin{aligned}&\mathbb{P}\left(\left|\max_{\boldsymbol{x}\in N}\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2-\rho nn_3\right|\geq c_3(\psi(\rho))^2nn_3\right)\\ \leq&2\cdot 5^n\cdot\exp\left(-c_2n\min\left(\left(\frac{c_3}{4c_1}\right)^2,\frac{c_3}{4c_1}\right)\right).\end{aligned}$$

Furthermore, taking the union bound over all $i=1,\cdots,n_3$, we have

$$\begin{aligned}&\mathbb{P}\left(\max_i\left|\max_{\boldsymbol{x}\in N}\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2-\rho nn_3\right|\geq c_3(\psi(\rho))^2nn_3\right)\\ \leq&2\cdot 5^n\cdot n_3\cdot\exp\left(-c_2n\min\left(\left(\frac{c_3}{4c_1}\right)^2,\frac{c_3}{4c_1}\right)\right).\end{aligned}$$

This implies that, with high probability (when the constant c_3 is large enough),

$$\max_i\max_{\boldsymbol{x}\in N}\|\bar{\boldsymbol{M}}_i\boldsymbol{x}\|_2^2\leq(\rho+c_3(\psi(\rho))^2)nn_3. \tag{71}$$

Let $\varphi(\rho)=2\sqrt{\rho+c_3(\psi(\rho))^2}$ and it satisfies $\lim_{\rho\to 0^+}\varphi(\rho)=0$ by using (70). The proof is completed by further combining (67), (68) and (71). □

E.2 Proof of Lemma D.2

Proof. For any tensor $\boldsymbol{\mathcal{Z}}$, we can write

$$\begin{aligned}&(\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}-\boldsymbol{\mathcal{P}}_{\boldsymbol{T}})\boldsymbol{\mathcal{Z}}\\ =&\sum_{ijk}\left(\rho^{-1}\delta_{ijk}-1\right)\langle\mathfrak{e}_{ijk},\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{Z}}\rangle\,\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\\ :=&\sum_{ijk}\boldsymbol{\mathcal{H}}_{ijk}(\boldsymbol{\mathcal{Z}})\end{aligned}$$

where $\boldsymbol{\mathcal{H}}_{ijk}:\mathbb{R}^{n\times n\times n_3}\to\mathbb{R}^{n\times n\times n_3}$ is a self-adjoint random operator with $\mathbb{E}[\boldsymbol{\mathcal{H}}_{ijk}]=\boldsymbol{0}$. Define the matrix operator $\bar{\boldsymbol{H}}_{ijk}:\mathbb{B}\to\mathbb{B}$, where $\mathbb{B}=\{\bar{\boldsymbol{B}}:\boldsymbol{\mathcal{B}}\in\mathbb{R}^{n\times n\times n_3}\}$ denotes the set consists of block diagonal matrices with the blocks as the frontal slices of $\bar{\boldsymbol{\mathcal{B}}}$, as

$$\bar{\boldsymbol{H}}_{ijk}(\bar{\boldsymbol{Z}})=\left(\rho^{-1}\delta_{ijk}-1\right)\langle\mathfrak{e}_{ijk},\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\rangle\,\mathtt{bdiag}(\overline{\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})}).$$

By the above definitions, we have $\|\boldsymbol{\mathcal{H}}_{ijk}\|=\|\bar{\boldsymbol{H}}_{ijk}\|$ and $\|\sum_{ijk}\boldsymbol{\mathcal{H}}_{ijk}\|=\|\sum_{ijk}\bar{\boldsymbol{H}}_{ijk}\|$. Also $\bar{\boldsymbol{H}}_{ijk}$ is self-adjoint and $\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}]=0$. To prove the result by the non-commutative Bernstein inequality, we need to bound $\|\bar{\boldsymbol{H}}_{ijk}\|$ and $\left\|\sum_{ijk}\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}^2]\right\|$. First, we have

$$\begin{aligned}\|\bar{\boldsymbol{H}}_{ijk}\|=&\sup_{\|\bar{\boldsymbol{Z}}\|_F=1}\|\bar{\boldsymbol{H}}_{ijk}(\bar{\boldsymbol{Z}})\|_F\\ \leq&\sup_{\|\bar{\boldsymbol{Z}}\|_F=1}\rho^{-1}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F\|\mathtt{bdiag}(\overline{\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})})\|_F\|\boldsymbol{\mathcal{Z}}\|_F\\ =&\sup_{\|\bar{\boldsymbol{Z}}\|_F=1}\rho^{-1}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\|\bar{\boldsymbol{Z}}\|_F\\ \leq&\frac{2\mu r}{nn_3\rho},\end{aligned}$$

where the last inequality uses (47). On the other hand, by direct computation, we have $\bar{\boldsymbol{H}}_{ijk}^2(\bar{\boldsymbol{Z}})=(\rho^{-1}\delta_{ijk}-$

$1)^2 \langle \mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\rangle \langle \mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\rangle \, \mathtt{bdiag}(\overline{\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})})$. Note that $\mathbb{E}[(\rho^{-1}\delta_{ijk}-1)^2] \le \rho^{-1}$. We have

$$
\begin{aligned}
&\left\|\sum_{ijk}\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}^2(\bar{\boldsymbol{Z}})]\right\|_F \\
\le&\rho^{-1}\left\|\sum_{ijk}\langle \mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\rangle \langle \mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\rangle \, \mathtt{bdiag}(\overline{\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})})\right\|_F \\
\le&\rho^{-1}\sqrt{n_3}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\left\|\sum_{ijk}\langle \mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\rangle\right\|_F \\
=&\rho^{-1}\sqrt{n_3}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\boldsymbol{\mathcal{Z}})\|_F \\
\le&\rho^{-1}\sqrt{n_3}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\|\boldsymbol{\mathcal{Z}}\|_F \\
=&\rho^{-1}\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F^2\|\bar{\boldsymbol{Z}}\|_F \\
\le&\frac{2\mu r}{nn_3\rho}\|\bar{\boldsymbol{Z}}\|_F.
\end{aligned}
$$

This implies $\left\|\sum_{ijk}\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}^2]\right\| \le \frac{2\mu r}{nn_3\rho}$. Let $\epsilon \le 1$. By Lemma E.1, we have

$$
\begin{aligned}
&\mathbb{P}\left[\|\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}-\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| > \epsilon\right] \\
=&\mathbb{P}\left[\left\|\sum_{ijk}\boldsymbol{\mathcal{H}}_{ijk}\right\| > \epsilon\right] \\
=&\mathbb{P}\left[\left\|\sum_{ijk}\bar{\boldsymbol{H}}_{ijk}\right\| > \epsilon\right] \\
\le&2nn_3\exp\left(-\frac{3}{8}\cdot\frac{\epsilon^2}{2\mu r/(nn_3\rho)}\right) \\
\le&2(nn_3)^{1-\frac{3}{16}C_0},
\end{aligned}
$$

where the last inequality uses $\rho \ge C_0\epsilon^{-2}\mu r\log(nn_3)/(nn_3)$. Thus, $\|\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}-\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \epsilon$ holds with high probability for some numerical constant C_0. □

E.3 Proof of Corollary D.3

Proof. From Lemma D.2, we have

$$\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}} - (1-\rho)^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \epsilon,$$

provided that $1-\rho \ge C_0\epsilon^{-2}(\mu r\log(nn_3))/n$. Note that $\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}} + \boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}$, we have

$$\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}-(1-\rho)^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}^\perp}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| = (1-\rho)^{-1}(\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}-\rho\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}).$$

Then, by the triangular inequality

$$\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| \le \epsilon(1-\rho)+\rho\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\| = \rho+\epsilon(1-\rho).$$

The proof is completed by using $\|\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\|^2 = \|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\|$. □

E.4 Proof of Lemma D.4

Proof. For any tensor $\boldsymbol{\mathcal{Z}} \in \boldsymbol{T}$, we write

$$\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}}) = \sum_{ijk}\rho^{-1}\delta_{ijk}z_{ijk}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk}).$$

The (a,b,c)-th entry of $\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}})-\boldsymbol{\mathcal{Z}}$ can be written as a sum of independent random variables, i.e.,

$$
\begin{aligned}
&\left\langle\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}})-\boldsymbol{\mathcal{Z}}, \mathfrak{e}_{abc}\right\rangle \\
=&\sum_{ijk}(\rho^{-1}\delta_{ijk}-1)z_{ijk}\left\langle\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk}), \mathfrak{e}_{abc}\right\rangle \\
:=&\sum_{ijk}t_{ijk},
\end{aligned}
$$

where t_{ijk}'s are independent and $\mathbb{E}(t_{ijk})=0$. Now we bound $|t_{ijk}|$ and $|\sum_{ijk}\mathbb{E}[t_{ijk}^2]|$. First

$$
\begin{aligned}
&|t_{ijk}| \\
\le&\rho^{-1}\|\boldsymbol{\mathcal{Z}}\|_\infty\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk})\|_F\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{abc})\|_F \\
\le&\frac{2\mu r}{nn_3\rho}\|\boldsymbol{\mathcal{Z}}\|_\infty.
\end{aligned}
$$

Second, we have

$$
\begin{aligned}
&\left|\sum_{ijk}\mathbb{E}[t_{ijk}^2]\right| \\
\le&\rho^{-1}\|\boldsymbol{\mathcal{Z}}\|_\infty^2\sum_{ijk}\left\langle\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{ijk}), \mathfrak{e}_{abc}\right\rangle^2 \\
=&\rho^{-1}\|\boldsymbol{\mathcal{Z}}\|_\infty^2\sum_{ijk}\left\langle\mathfrak{e}_{ijk}, \boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{abc})\right\rangle^2 \\
=&\rho^{-1}\|\boldsymbol{\mathcal{Z}}\|_\infty^2\|\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}(\mathfrak{e}_{abc})\|_F^2 \\
\le&\frac{2\mu r}{nn_3\rho}\|\boldsymbol{\mathcal{Z}}\|_\infty^2.
\end{aligned}
$$

Let $\epsilon \le 1$. By Lemma E.1, we have

$$
\begin{aligned}
&\mathbb{P}\left[|[\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}})-\boldsymbol{\mathcal{Z}}]_{abc}| > \epsilon\|\boldsymbol{\mathcal{Z}}\|_\infty\right] \\
=&\mathbb{P}\left[\left|\sum_{ijk}t_{ijk}\right| > \epsilon\|\boldsymbol{\mathcal{Z}}\|_\infty\right] \\
\le&2\exp\left(-\frac{3}{8}\cdot\frac{\epsilon^2\|\boldsymbol{\mathcal{Z}}\|_\infty^2}{2\mu r\|\boldsymbol{\mathcal{Z}}\|_\infty^2/(nn_3\rho)}\right) \\
\le&2(nn_3)^{-\frac{3}{16}C_0},
\end{aligned}
$$

where the last inequality uses $\rho \ge C_0\epsilon^{-2}\mu r\log(nn_3)/(nn_3)$. Thus, $\|\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{T}}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{Z}})-\boldsymbol{\mathcal{Z}}\|_\infty \le \epsilon\|\boldsymbol{\mathcal{Z}}\|_\infty$ holds with high probability for some numerical constant C_0. □

E.5 Proof of Lemma D.5

Proof. Denote the tensor $\boldsymbol{\mathcal{H}}_{ijk} = \left(1-\rho^{-1}\delta_{ijk}\right)z_{ijk}\mathfrak{e}_{ijk}$. Then we have

$$(\boldsymbol{\mathcal{I}}-\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})\boldsymbol{\mathcal{Z}} = \sum_{ijk}\boldsymbol{\mathcal{H}}_{ijk}.$$

Note that δ_{ijk}'s are independent random scalars. Thus, $\boldsymbol{\mathcal{H}}_{ijk}$'s are independent random tensors and $\bar{\boldsymbol{H}}_{ijk}$'s are independent random

matrices. Observe that $\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}] = \mathbf{0}$ and $\|\bar{\boldsymbol{H}}_{ijk}\| \leq \rho^{-1}\|\boldsymbol{\mathcal{Z}}\|_\infty$. We have

$$\begin{aligned}
&\left\|\sum_{ijk}\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}^*\bar{\boldsymbol{H}}_{ijk}]\right\| \\
=&\left\|\sum_{ijk}\mathbb{E}[\boldsymbol{\mathcal{H}}_{ijk}^**\boldsymbol{\mathcal{H}}_{ijk}]\right\| \\
=&\left\|\sum_{ijk}\mathbb{E}[(1-\rho^{-1}\delta_{ijk})^2]z_{ijk}^2(\mathring{\mathfrak{e}}_j*\mathring{\mathfrak{e}}_j^*)\right\| \\
=&\left\|\frac{1-\rho}{\rho}\sum_{ijk}z_{ijk}^2(\mathring{\mathfrak{e}}_j*\mathring{\mathfrak{e}}_j^*)\right\| \\
\leq&\frac{nn_3}{\rho}\|\boldsymbol{\mathcal{Z}}\|_\infty^2.
\end{aligned}$$

A similar calculation yields $\left\|\sum_{ijk}\mathbb{E}[\bar{\boldsymbol{H}}_{ijk}^*\bar{\boldsymbol{H}}_{ijk}]\right\| \leq \rho^{-1}nn_3\|\boldsymbol{\mathcal{Z}}\|_\infty^2$. Let $t = \sqrt{C_0nn_3\log(nn_3)/\rho}\|\boldsymbol{\mathcal{Z}}\|_\infty$. When $\rho \geq C_0\log(nn_3)/(nn_3)$, we apply Lemma E.1 and obtain

$$\begin{aligned}
&\mathbb{P}\left[\|(\boldsymbol{\mathcal{I}}-\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})\boldsymbol{\mathcal{Z}}\| > t\right] \\
=&\mathbb{P}\left[\left\|\sum_{ijk}\boldsymbol{\mathcal{H}}_{ijk}\right\| > t\right] \\
=&\mathbb{P}\left[\left\|\sum_{ijk}\bar{\boldsymbol{H}}_{ijk}\right\| > t\right] \\
\leq&2nn_3\exp\left(-\frac{3}{8}\cdot\frac{C_0nn_3\log(nn_3)\|\boldsymbol{\mathcal{Z}}\|_\infty^2/\rho}{nn_3\|\boldsymbol{\mathcal{Z}}\|_\infty^2/\rho}\right) \\
\leq&2(nn_3)^{1-\frac{3}{8}C_0}.
\end{aligned}$$

Thus, $\|(\boldsymbol{\mathcal{I}}-\rho^{-1}\boldsymbol{\mathcal{P}}_{\boldsymbol{\Omega}})\boldsymbol{\mathcal{Z}}\| > t$ holds with high probability for some numerical constant C_0. □