Tensor Robust Principal Component Analysis from Multi-Level Quantized Observations

Jianjun Wang, *Member, IEEE,* Jingyao Hou, and Yonina C. Eldar, *Fellow, IEEE*

***Abstract*—We consider Quantized Tensor Robust Principal Component Analysis (Q-TRPCA), which aims to recover a low-rank tensor and a sparse tensor from noisy, quantized, and sparsely corrupted measurements. A nonconvex constrained maximum likelihood (ML) estimation method is proposed for Q-TRPCA. We provide an upper bound on the Frobenius norm of tensor estimation error under this method. Making use of tools in information theory, we derive a theoretical lower bound on the best achievable estimation error from unquantized measurements. Compared with the lower bound, the upper bound on the estimation error is nearly order-optimal. We further develop an efficient convex ML estimation scheme for Q-TRPCA based on the tensor nuclear norm (TNN) constraint. This method is more robust to sparse noises than the latter nonconvex ML estimation approach. Conducting experiments on both synthetic data and real-world data, we show the effectiveness of the proposed methods.**

***Index Terms*—Low-tubal-rank tensor, multi-level quantization, maximum likelihood estimation, proximal algorithm.**

I. Introduction

PRINCIPAL component analysis (PCA) is an important and classical data analysis approach, which aims to reduce the dimensionality of a high-dimensional data set. It has been widely utilized in various applications, such as chemometrics, hyperspectral image recovery, computer vision, and more. However, a major issue of PCA is that it has difficulty handling large noises and severe outliers that are ubiquitous in actual data. To overcome this shortcoming, robust PCA (RPCA) [1] has been proposed. Consider an observed matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, which can be decomposed as $\boldsymbol{X} = \boldsymbol{L}_0 + \boldsymbol{C}_0$, where $\boldsymbol{L}_0$ is a low-rank matrix and $\boldsymbol{C}_0$ is a sparse matrix. RPCA aims to recover $\boldsymbol{L}_0$ and $\boldsymbol{C}_0$ from the observation $\boldsymbol{X}$. This method has been well studied and successfully applied to many fields, including foreground-background separation [1], video compressive sensing [2], latent semantic indexing [3], ultrasound imaging [4], and more.

Practical datasets typically consist of quantized measurements where quantization with a small number of bits allows to efficiently compress large-scale data to alleviate challenges in terms of data storage and transfer. Furthermore, quantization combined with noise can effectively enhance data privacy [5]–[7]. Suppose that we are given quantized measurements $\boldsymbol{Y} = Q(\boldsymbol{L}_0 + \boldsymbol{C}_0)$, where $\boldsymbol{L}_0$ denotes a low-rank matrix, $\boldsymbol{C}_0$ is a sparse matrix, and Q is a K-level quantization operator which applies componentwise on $\boldsymbol{L}_0 + \boldsymbol{C}_0$. When K is large, one could treat quantization error simply as bounded noise and use methods for RPCA. When K is small, this approach is not satisfactory. In fact, the recovery problem is often ill-posed in this case [8] since low-level quantization loses a lot of the information on the entries. To make the recovery problem well-posed, quantized RPCA (Q-RPCA) has been considered [7], [9], which adds random dithers prior to quantization so that $\boldsymbol{Y} = Q(\boldsymbol{L}_0 + \boldsymbol{C}_0 + \boldsymbol{N})$ where $\boldsymbol{N}$ is a dither matrix consisting of i.i.d. random entries generated from a known distribution. Q-RPCA aims to recover the low-rank matrix $\boldsymbol{L}_0$ and the sparse matrix $\boldsymbol{C}_0$ from $\boldsymbol{Y}$ under certain conditions. Note that the dithers are unknown in the recovery process and we only know the distribution generating them.

The problem of Q-RPCA falls within recent research on low-rank matrix recovery from quantized measurements [7]–[17]. The common feature of the existing recovery methods is to relate the quantized measurements with the underlying needed-to-be recovered low-rank matrix $\boldsymbol{L}_0$ with a probabilistic model and estimate $\boldsymbol{L}_0$ by solving a constrained maximum likelihood (ML) optimization problem. However, most of the existing works focus on matrix completion that recovers the underlying needed-to-be recovered low-rank matrix from incomplete quantized observations, which significantly differs from Q-RPCA in both the observation model and recovery task. To solve Q-RPCA, Lan et al. [9] proposed an ML optimization problem with a nuclear norm constraint as a low-rankness promotion and an l_1-norm constraint as a sparsity promotion, which can be addressed by a multi-convex algorithm. Gao et al. [7] considered an ML optimization problem with tighter constraints, i.e., a bounded rank constraint for the low-rank component and a constraint on the number of nonzero entries for the sparse component. They also provided an algorithm based on matrix factorization to solve their program problem.

One major drawback of Q-RPCA is that it can only effectively handle two-order (matrix) data. However, in many cases such as color image processing [18], hyperspectral image recovery [19], video processing [20], and context-aware recommendation systems [21], the data is multi-dimensional. Such multi-dimensional arrays are known as tensors [22]. For example, a color image is a three-order tensor with column, row, and color modes; greyscale video data can be represented

This work was supported by National Natural Science Foundation of China (Grant Nos. 12071380, 61673015).

Jianjun Wang is with the School of Mathematics and Statistics, Southwest University, Chongqing 400715, China (e-mail: wjjmath@163.com; wjj@swu.edu.cn).

Jingyao Hou is with the College of Mathematics & Information, China West Normal University, Nanchong 637002, Sichuan, China (e-mail: hjy17623226280@163.com; houjingyao@email.swu.edu.cn).

Yonina. C. Eldar is with the Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

(*Corresponding author: Jingyao Hou.)

as {rows of a frame $\times$ columns of a frame $\times$ different frames}. To use Q-RPCA for a tensor, one has to first transform the multi-dimensional array into matrix form. Such preprocessing destroys the intrinsic tensor structure of multi-dimensional data. For instance, if each frame of a video is vectorized to construct a matrix, the spatial correlation will no longer be preserved. To alleviate this issue, it is natural to extend the framework of Q-RPCA to directly handle tensors.

The tensor extension of existing methods for RPCA is not easy. The main issue is that the tensor rank is not well defined. There are several tensor rank definitions and corresponding convex relaxation functions, but each has its limitation. For example, the CANDECOMP/PARAFAC (CP) rank [22] is defined by the smallest number of rank-one tensors in the CP decomposition. However, the calculation of CP rank is generally an NP-hard problem. There are several convex relaxation functions for the CP rank such as nuclear norm (see e.g. [23] and [24]) and atomic M-norm [25]. However, calculating the nuclear norm of a tensor is still NP-hard [26], and how to design a polynomial-time method for the atomic M-norm constrained optimization problem is also not clear [25]. Another example is the tractable Tucker rank [27], which is defined as $\mathrm{rank_{tc}}(\boldsymbol{\mathcal{Z}}) = \{\mathrm{rank}(\boldsymbol{Z}_{(1)}), \mathrm{rank}(\boldsymbol{Z}_{(2)}), \cdots, \mathrm{rank}(\boldsymbol{Z}_{(l)})\}$, where $\boldsymbol{Z}_{(i)}$ is the mode-i matricization of $\boldsymbol{\mathcal{Z}}$. However, the Tucker rank cannot appropriately capture the global correlation of a tensor. The main reason is that only a single mode represents the matrix row in $\boldsymbol{Z}_{(i)}$ which is an unbalanced matricization scheme [28]. In addition, as a common convex surrogate of the Tucker rank, the Sum of Nuclear Norms (SNN) [29] is not the tightest convex relaxation.

Recently, a new tensor decomposition scheme, tensor Singular Value Decomposition (t-SVD), was proposed [30]. The t-SVD enjoys several similar properties to the matrix SVD. Following in this vein, the tubal rank [31] and its convex surrogate, tensor nuclear norm [32], [33], were proposed to characterize the inherent structure of tensor data. Many tensor data arising in applications have been shown to have (approximately) low-tubal-rank structures [20], [34], [35]. Furthermore, Lu et al. [34] defined the tensor average rank and proved that a tensor always has a low average rank if it has a low tubal rank or a low CP rank. They also define a new tensor nuclear norm and show that it is a convex envelope of the tensor average rank within the unit ball of the tensor spectral norm. Therefore, t-SVD overcomes the limitations of other decomposition methods to some extent. Consequently, tensor robust PCA (TRPCA) has been considered under the framework of t-SVD as an extension of RPCA to tensor data [34].

We consider quantized tensor robust PCA (Q-TRPCA) under the framework of t-SVD in this paper. As an extension of Q-RPCA, Q-TRPCA aims to recover a low-rank tensor from noisy, quantized, and erroneous measurements. More specifically, consider an underlying tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ which can be decomposed as $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{C}}_0$ where $\boldsymbol{\mathcal{L}}_0$ has a low tubal rank and $\boldsymbol{\mathcal{C}}_0$ is sparse. Suppose that we are given $\boldsymbol{\mathcal{Y}} = Q(\boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{C}}_0 + \boldsymbol{\mathcal{N}})$, where $\boldsymbol{\mathcal{N}}$ is a random dither tensor that consists of i.i.d. entries generated from a known cumulative distribution function $g(\cdot)$, and $Q(\cdot)$ is a K-level quantization operator that applies componentwise on $\boldsymbol{\mathcal{L}}_0 + \boldsymbol{\mathcal{C}}_0 + \boldsymbol{\mathcal{N}}$. Our goal is to recover the low-tubal-rank tensor $\boldsymbol{\mathcal{L}}_0$ and the sparse tensor $\boldsymbol{\mathcal{C}}_0$ from observations $\boldsymbol{\mathcal{Y}}$ given $g(\cdot)$ and $Q(\cdot)$. In this paper, we formally analyze this problem and find the estimates of $\boldsymbol{\mathcal{L}}_0$ and $\boldsymbol{\mathcal{C}}_0$ by maximizing log-likelihood estimation functions given observations $\boldsymbol{\mathcal{Y}}$. The main contributions of this paper are as follows.

- We propose a nonconvex ML optimization problem for Q-TRPCA, followed by a theoretical analysis of the error between the underlying needed-to-be recovered tensor and the estimate obtained by solving the proposed problem. We also derive a theoretical lower bound on the best achievable estimation error from unquantized measurements. Compared with the lower bound, the estimation error by solving the problem is nearly order-optimal.
- We suggest a convex surrogate for the proposed nonconvex problem, which is more robust to sparse noises. Solving this convex optimization problem usually obtains better performance in practice.
- Experiments on synthetic and real-world data show the validity of the proposed methods. In particular, we apply our methods to color image and multispectral image recovery from highly quantized and sparsely corrupted observations, which highlights the practical aspects of our schemes.

As for the problem of low-rank plus sparse tensor estimation without quantization, several recent papers have provided strong theoretical results. For example, Zhang et al. [36] developed comprehensive results on both the statistical and computational limits for the problem of extracting the underlying Tucker low-rank structure from the noisy tensor observations, exhibiting three distinct phases according to the signal-to-noise ratio (SNR); Cai et al. [37] proposed a fast algorithm for the problem by integrating the Riemannian gradient descent and a novel gradient pruning procedure which is shown to converge linearly under suitable conditions; Han et al. [38] described a flexible framework for generalized low-rank tensor estimation problems that includes sub-Gaussian tensor PCA, tensor regression, and Poisson and binomial tensor PCA and proposed an algorithm achieving the minimax optimal rate of convergence in estimation error. As for the case with quantization, Aidini et al. [39] solved the problem through matricization in each mode of the tensor and then utilize methods for Q-RPCA (thus we name the procedure T2M-QRPCA for abbreviation). The recovered tensor is then constructed as the weighted sum of the recovered unfolded matrices. In our work, on the other hand, we recover the underlying low-rank tensor under t-SVD so that the spatial correlation of tensor data is better maintained. Q-TRPCA also relates to several lines of existing works on recovering a low-rank tensor from quantized and incomplete measurements [40]–[43], [45]. Refs. [40]–[43] provide recovery guarantees when the measurements are binary and incomplete. Ref. [45] considers the problem of low-rank tensor estimation from possibly incomplete, ordinal-valued observations. These works do not consider measurements corrupted by sparse noises, which significantly differs from the problem formulation of

Q-TRPCA.

The rest of this paper is structured as follows. Section II clarifies some notations and provides an overview of the tensor algebra framework under t-SVD. Section III reviews results for matrix quantized RPCA and introduces the problem formulations we are interested in. In Section IV, we propose two recovery schemes and derive corresponding theoretical guarantees. Section V introduces specific algorithms and Section VI presents numerical experiments based on synthetic and real-world data. Finally, we conclude the paper in Section VII. In Appendix A, we prove our main theoretical results for the proposed recovery schemes.

II. Overview of Tensors Algebraic Framework

A. Notations

We begin with some basic notations used throughout the paper. Scalars are denoted with lowercase letters, e.g., a, vectors with boldface lowercase letters, e.g., $\boldsymbol{a}$, matrices with boldface capital letters, e.g., $\boldsymbol{A}$, and tensors with boldface Euler script letters, e.g., $\boldsymbol{\mathcal{A}}$. The fields of real numbers and complex numbers are denoted as $\mathbb{R}$ and $\mathbb{C}$, respectively. For a finite set $\mathbb{A}$, $|\mathbb{A}|$ is the number of elements. Let $[n] := \{1, 2, \cdots, n\}$ be the set consisting of the nonzero integers not greater than n. For a three-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$, we write its (i, j, k)-th entry as $\mathcal{A}_{ijk}$ and use Matlab notations $\boldsymbol{\mathcal{A}}(i, :, :)$, $\boldsymbol{\mathcal{A}}(:, i, :)$, and $\boldsymbol{\mathcal{A}}(:, :, i)$ to denote respectively the i-th horizontal, lateral, and frontal slice. The i-th frontal slice is also written as $\boldsymbol{A}^{(i)}$ for simplicity. The tubes of $\boldsymbol{\mathcal{A}}$ are denoted as $\boldsymbol{\mathcal{A}}(i, j, :)$. The inner product between matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n_1 \times n_2}$ is defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}(\boldsymbol{A}^* \boldsymbol{B})$, where $\boldsymbol{A}^*$ is the conjugate transpose of $\boldsymbol{A}$ and $\mathrm{Tr}(\cdot)$ is the matrix trace. The inner product between tensors $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ of the same size is defined as $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle_F = \sum_k \langle \boldsymbol{A}^{(k)}, \boldsymbol{B}^{(k)} \rangle_F = \sum_{i,j,k} \mathcal{A}_{ijk} \mathcal{B}_{ijk}$. Suppose that $g(m)$ and $h(m)$ are functions of m; $h(m)$ is on the order $\Theta(g(m))$ means that $c_2 \cdot g(m) \leq h(m) \leq c_1 \cdot g(m)$ holds for some positive constants c_1 and c_2 as m goes to infinity. We say that $h(m)$ is on the order $\mathcal{O}(g(m))$ when $h(m) \leq c_1 \cdot g(m)$ holds for some positive constant c_1 as m goes to infinity. We introduce the operator $\mathrm{vec}(\boldsymbol{\mathcal{A}})$ as the vectorization of the tensor $\boldsymbol{\mathcal{A}}$. For simplicity, we always set $n_{(1)} = \max\{n_1, n_2\}$ and $n_{(2)} = \min\{n_1, n_2\}$.

We denote the infinity norm as $\|\boldsymbol{\mathcal{A}}\|_\infty = \max_{ijk} |\mathcal{A}_{ijk}|$ and the Frobenius norm as $\|\boldsymbol{\mathcal{A}}\|_F = \sqrt{\sum_{ijk} |\mathcal{A}_{ijk}|^2}$, respectively. The above norms reduce to the vector or the matrix norms if $\boldsymbol{\mathcal{A}}$ is a vector or a matrix. Following the notations in compressed sensing, with some abuse, we set $\|\boldsymbol{\mathcal{A}}\|_0 = \sum_{ijk} \mathbb{1}_{\{\mathcal{A}_{ijk} \neq 0\}}$ to denote the number of nonzero entries in $\boldsymbol{\mathcal{A}}$, where $\mathbb{1}_{[\boldsymbol{\varepsilon}]}$ is the indicator function for the event $\boldsymbol{\varepsilon}$, i.e., $\mathbb{1}_{\{\boldsymbol{\varepsilon}\}}$ is 1 if $\boldsymbol{\varepsilon}$ occurs and 0 otherwise. We also set $\|\boldsymbol{\mathcal{A}}\|_1 = \sum_{ijk} |\mathcal{A}_{ijk}|$. For a matrix $\boldsymbol{A} \in \mathbb{C}^{n_1 \times n_2}$, the spectral norm is $\|\boldsymbol{A}\| = \max_i \sigma_i(\boldsymbol{A})$, where $\sigma_i(\boldsymbol{A})$'s are the singular values of $\boldsymbol{A}$. The matrix nuclear norm of $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A})$.

For a three-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote $\bar{\boldsymbol{\mathcal{A}}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ as the Discrete Fourier Transform (DFT) along the third dimension of $\boldsymbol{\mathcal{A}}$, i.e., performing the DFT on all tubes of $\boldsymbol{\mathcal{A}}$. We also define the block diagonal matrix

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

whose i-th block on the diagonal is the i-th frontal slice of $\bar{\boldsymbol{\mathcal{A}}}$. Here $\mathrm{bdiag}(\cdot)$ is a map transforming a tensor into a block diagonal matrix according to frontal slices. We define $\mathrm{bcirc}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{n_1 n_3 \times n_2 n_3}$ as the block circulant matrix of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, i.e.,

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

Two operators, unfold$(\cdot)$ and fold$(\cdot)$, are introduced as follows

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

B. T-poduct and T-SVD

Next, we introduce background results in the framework of t-SVD. We first recall the definition of tensor-tensor product (t-product) between two three-order tensors.

Definition 1 (t-product, [30])**.** *For two tensors $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_2 \times l \times n_3}$, the t-product between $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ denoted as $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}$ is defined to be an $n_1 \times l \times n_3$ tensor $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} = \mathrm{fold}(\mathrm{bcirc}(\boldsymbol{\mathcal{A}}) \cdot \mathrm{unfold}(\boldsymbol{\mathcal{B}}))$.*

We also need a few more definitions before introducing the t-SVD.

Definition 2 (See [30])**.** *We introduce the following definitions.*

(i) *For a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, its transpose $\boldsymbol{\mathcal{A}}^T$ is an $n_2 \times n_1 \times n_3$ tensor obtained by transposing each of the frontal slices and then reversing the order of the transposed frontal slices 2 through n_3;*

(ii) *The identity tensor $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{n \times n \times n_3}$ denotes the tensor whose first frontal slice is the $n \times n$ identity matrix, and other frontal slices are all zero entries;*

(iii) *Tensor $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{n \times n \times n_3}$ is orthogonal if it satisfies $\boldsymbol{\mathcal{Q}}^T * \boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{Q}} * \boldsymbol{\mathcal{Q}}^T = \boldsymbol{\mathcal{I}}$;*

(iv) *A tensor is called F-diagonal if each of its frontal slices is a diagonal matrix.*

Based on these definitions, the t-SVD is defined as follows.

Proposition 1 (t-SVD, [30])**.** *Any three-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ can be factorized as $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^T$, 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 both orthogonal tensors, and $\boldsymbol{\mathcal{S}}$ is an F-diagonal tensor. The first frontal slice $\boldsymbol{\mathcal{S}}(:, :, 1)$ of $\boldsymbol{\mathcal{S}}$ satisfies $\mathcal{S}_{111} \geq \mathcal{S}_{221} \geq \cdots \geq \mathcal{S}_{n_{(2)} n_{(2)} 1}$.*

Under the framework of t-SVD, tensor tubal rank is defined which plays a similar role to matrix rank.

Definition 3 (Tubal rank, [31]). *The tensor tubal rank of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, denoted as $\mathrm{rank_t}(\boldsymbol{\mathcal{A}})$, is defined as the number of nonzero singular tubes of $\boldsymbol{\mathcal{S}}$, where $\boldsymbol{\mathcal{S}}$ is from the t-SVD of $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^T$. Due to the property of the DFT, we have*

$$\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) = \#\{i : \boldsymbol{\mathcal{S}}(i,i,:) \neq \mathbf{0}\} = \#\{i : \mathcal{S}_{ii1} \neq 0\}.$$

Proposition 2 (See [46]). *Any tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ with $\mathrm{rank_t}(\boldsymbol{\mathcal{A}}) = r$ can be written in a t-product form $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{F}} * \boldsymbol{\mathcal{H}}$, where $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{n_1\times r\times n_3}$ and $\boldsymbol{\mathcal{H}} \in \mathbb{R}^{r\times n_2\times n_3}$ are two tensors of smaller size and satisfy $\mathrm{rank_t}(\boldsymbol{\mathcal{F}}) = \mathrm{rank_t}(\boldsymbol{\mathcal{H}}) = r$.*

The spectral norm and nuclear norm of three-order tensors under the framework of t-SVD are defined as follows.

Definition 4 (Tensor spectral norm, [34]). *The tensor spectral norm of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, denoted as $\|\boldsymbol{\mathcal{A}}\|$, is defined as $\|\boldsymbol{\mathcal{A}}\| := \|\mathrm{bcirc}(\boldsymbol{\mathcal{A}})\| = \sup_{\|\boldsymbol{\mathcal{V}}\|_F\leq 1} \|\mathrm{bcirc}(\boldsymbol{\mathcal{A}})\cdot \mathrm{unfold}(\boldsymbol{\mathcal{V}})\|_F$, where $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2\times 1\times n_3}$.*

Definition 5 (Tensor nuclear norm (TNN), [34]). *Let $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^T$ 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 $\boldsymbol{\mathcal{S}}(:,:,1)$ are called the singular values of $\boldsymbol{\mathcal{A}}$. Thus the tensor nuclear norm of $\boldsymbol{\mathcal{A}}$ is denoted as $\|\boldsymbol{\mathcal{A}}\|_* := \sum_{i=1}^{r} \mathcal{S}_{ii1}$ with $r = \mathrm{rank_t}(\boldsymbol{\mathcal{A}})$.*

TNN and tensor Frobenius norm under t-SVD have the following properties:

Proposition 3 (See [34]). *For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, we have*

(i) $\|\boldsymbol{\mathcal{A}}\|_*$ *is the dual norm of* $\|\boldsymbol{\mathcal{A}}\|$;

(ii) $\|\boldsymbol{\mathcal{A}}\|_* = \dfrac{1}{n_3}\|\bar{\boldsymbol{A}}\|_* = \dfrac{1}{n_3}\|\mathrm{bcirc}(\boldsymbol{\mathcal{A}})\|_*$;

(iii) $\|\boldsymbol{\mathcal{A}}\|_F = \dfrac{1}{\sqrt{n_3}}\|\bar{\boldsymbol{A}}\|_F = \dfrac{1}{\sqrt{n_3}}\|\mathrm{bcirc}(\boldsymbol{\mathcal{A}})\|_F$.

III. Quantized Robust PCA

This section begins with reviewing quantized robust PCA for matrices and then introduces the problem formulation considered in this paper.

A. Quantized Robust PCA

Let $Q(\cdot)$ be a K-level quantization operator with quantization boundaries $\omega_0 \leq \omega_1 \leq \cdots < \omega_K$, which is defined by

$$Q(x) = l \text{ if } \omega_{l-1} < x \leq \omega_l,\ \forall l \in [K]. \tag{3}$$

In the Q-RPCA problem, measurements are in the form of

$$Y_{ij} = Q(\widetilde{L}_{ij} + \widetilde{C}_{ij} + N_{ij}),\ \forall (i,j) \in [n_1]\times[n_2], \tag{4}$$

where $\widetilde{\boldsymbol{L}} = (\widetilde{L}_{ij}) \in \mathbb{R}^{n_1\times n_2}$ denotes the actual data matrix whose rank does not exceed r, $\widetilde{\boldsymbol{C}} = (\widetilde{C}_{ij}) \in \mathbb{R}^{n_1\times n_2}$ is a sparse matrix that has at most s nonzero entries, $\boldsymbol{N} = (N_{ij}) \in \mathbb{R}^{n_1\times n_2}$ is the dither matrix that consists of i.i.d. entries generated from a cumulative distribution function $g(\cdot)$, and $Q(\cdot)$ is a K-level quantization operator.

Define $\widetilde{\boldsymbol{X}} := \widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{C}}$ and the following probability function

$$f_l(x) := g(\omega_l - x) - g(\omega_{l-1} - x), \tag{5}$$

for each $l = 0, 1, \cdots, K$. Then (4) implies that

$$Y_{ij} = l \text{ with probability } f_l(\widetilde{X}_{ij}),\ \forall (i,j) \in [n_1]\times[n_2],$$

where $\sum_{l=1}^{K} f_l(\widetilde{X}_{ij}) = 1$. Using the maximum log-likelihood estimation criterion, Lan et al. [9] proposed the following objective in order to recover $\widetilde{\boldsymbol{L}}$ and $\widetilde{\boldsymbol{C}}$:

$$\begin{aligned}(\hat{\boldsymbol{L}}, \hat{\boldsymbol{C}}) = \arg\min_{\boldsymbol{L},\boldsymbol{C}} &- \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{K} \mathbb{1}_{\{Y_{ij}=l\}} \log(f_l(L_{ij}+C_{ij})) \\ &\text{s.t. } \|\boldsymbol{L}\|_* \leq \lambda \text{ and } \|\boldsymbol{C}\|_1 \leq \mu,\end{aligned} \tag{6}$$

where the nuclear norm constraint $\|\boldsymbol{L}\|_* \leq \lambda$ promotes the low-rankness of $\boldsymbol{L}$ and $\|\boldsymbol{C}\|_1 \leq \mu$ corresponds to an ℓ_1-norm constraint on the vectorized version of $\boldsymbol{C}$ that promotes sparsity among all its entries. A block-coordinate descent (BCD) based algorithm was proposed in [9] to solve (6), which we name Q-RPCA(BCD) for abbreviation.

Ref. [9] did not provide theoretical analysis for their method. Gao et al. [7] consider a nonconvex problem

$$\begin{aligned}(\hat{\boldsymbol{L}}, \hat{\boldsymbol{C}}) = \arg\min_{\boldsymbol{L},\boldsymbol{C}} &- \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{K} \mathbb{1}_{\{Y_{ij}=l\}} \log(f_l(L_{ij}+C_{ij})) \\ &\text{s.t. } (\boldsymbol{L},\boldsymbol{C}) \in \mathbb{S}_{r,s},\end{aligned} \tag{7}$$

where $\mathbb{S}_{r,s}$ is defined as

$$\begin{aligned}\mathbb{S}_{r,s} := \big\{(\boldsymbol{L},\boldsymbol{C}) : \boldsymbol{L},\boldsymbol{C} \in \mathbb{R}^{n_1\times n_2}, \|\boldsymbol{L}\|_\infty \leq \alpha, \|\boldsymbol{C}\|_\infty \leq \alpha,\\ \mathrm{rank}(\boldsymbol{L}) \leq r, \|\boldsymbol{C}\|_0 \leq s\big\}.\end{aligned}$$

Note that the assumptions $\|\boldsymbol{L}\|_\infty \leq \alpha$, $\|\boldsymbol{C}\|_\infty \leq \alpha$ in $\mathbb{S}_{r,s}$ are used to facilitate the establishment of recovery guarantees. In RPCA, the incoherence assumption is used to impose that the low-rank component $\boldsymbol{L}$ is not sparse. The assumptions $\|\boldsymbol{L}\|_\infty \leq \alpha$, $\|\boldsymbol{C}\|_\infty \leq \alpha$ are less stringent than the incoherence assumption [12], which help make the recovery of $\widetilde{\boldsymbol{L}}$ and $\widetilde{\boldsymbol{C}}$ well-posed by preventing excessive "spikiness" of the matrices. We refer the reader to [8], [12], [14]–[16], [47] for further details. In addition, some assumptions on f_l in (7) are needed for the recovery to facilitate analysis. Gao et al. define constants γ_α and L_α as

$$\gamma_\alpha = \min_{l\in[K]} \inf_{|x|\leq 2\alpha} \left\{ \frac{\dot{f}_l^2(x)}{f_l^2(x)} - \frac{\ddot{f}_l(x)}{f_l(x)} \right\}, \tag{8}$$

$$L_\alpha = \max_{l\in[K]} \sup_{|x|\leq 2\alpha} \left\{ |\dot{f}_l(x)|/f_l(x) \right\}, \tag{9}$$

where $\dot{f}_l(x)$ and $\ddot{f}_l(x)$ denote the first and second order derivatives of f_l with respect to x. Note that the values of γ_α and L_α rely on α, g, and ω_i's ($i \in \{0, 1, \cdots, K\}$) but are independent of n_1, n_2, and n_3. It is known that $f_l(x)$ is log-concave if and only if $(\dot{f}_l(x))^2 \geq \ddot{f}_l(x)\dot{f}_l(x)$ [52]. Thus $\gamma_\alpha \geq 0$ for log-concave f_l and $\gamma_\alpha > 0$ for strictly log-concave f_l. In particular, when random noises are generated from the logistic distribution $g(x) := \Phi_{log}(\frac{x}{\sigma}) = 1/(1+e^{-x/\sigma})$ and the Gaussian distribution $g(x) := \Phi_{norm}(\frac{x}{\sigma})$, one can check that

f_l's are strictly log-concave, and thus $\gamma_\alpha > 0$. For the logistic distribution, it turns out that $L_\alpha = 1/(2\sigma\beta_{\alpha\sigma})$ [12] where

$$\beta_{\alpha\sigma} := \min_{l\in[K]} \inf_{|x|\le\alpha} \left\{ \Phi_{log}(\frac{\omega_l - x}{\sigma}) - \Phi_{log}(\frac{\omega_{l-1} - x}{\sigma}) \right\}.$$

For the Gaussian distribution, $L_\alpha = \sqrt{2}/(\sqrt{\pi}\sigma\beta'_{\alpha\sigma})$ [12] where

$$\beta'_{\alpha\sigma} := \min_{l\in[K]} \inf_{|x|\le\alpha} \left\{ \Phi_{norm}(\frac{\omega_l - x}{\sigma}) - \Phi_{norm}(\frac{\omega_{l-1} - x}{\sigma}) \right\}.$$

With assumptions (8) and (9), for strictly log-concave f_l's, Gao et al. [7] proved that

$$\frac{\|(\hat{\boldsymbol{L}} + \hat{\boldsymbol{C}}) - (\widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{C}})\|_F}{\sqrt{n_1 n_2}} \le \min\left(2\alpha\left(1 + \sqrt{\frac{2s}{n_1 n_2}}\right), U'_\alpha\right), \tag{10}$$

where

$$U'_\alpha = \max\left(\frac{8.04 L_\alpha \sqrt{2r}}{\gamma_\alpha \sqrt{n_{(2)}}}, 4\sqrt{\frac{2.01\alpha L_\alpha \sqrt{r n_{(1)} s} + \alpha L_\alpha s}{\gamma_\alpha n_1 n_2}}\right)$$

and furthermore

$$\frac{\|\hat{\boldsymbol{L}} - \widetilde{\boldsymbol{L}}\|_F}{\sqrt{n_1 n_2}} \le \min\left(2\alpha\left(1 + 2\sqrt{\frac{2s}{n_1 n_2}}\right), U'_\alpha + 2\alpha\sqrt{\frac{2s}{n_1 n_2}}\right) \tag{11}$$

with a high probability. As long as s is at most $\Theta(n_{(1)})$, the above results imply that

$$\|(\hat{\boldsymbol{L}} + \hat{\boldsymbol{C}}) - (\widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{C}})\|_F / \sqrt{n_1 n_2} \le \mathcal{O}\left(\sqrt{2r/n_{(2)}}\right) \tag{12}$$

and

$$\|\hat{\boldsymbol{L}} - \widetilde{\boldsymbol{L}}\|_F / \sqrt{n_1 n_2} \le \mathcal{O}\left(\sqrt{2r/n_{(2)}}\right). \tag{13}$$

Gao et al. also showed that the minimum possible recovery error from unquantized noisy measurements $\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{C} + \boldsymbol{N}$ is at least on the order of $\mathcal{O}(\sqrt{r/n_{(2)}})$. In that sense, the estimate obtained by solving (7) is order-optimal when s is at most $\Theta(n_{(1)})$. A proximal alternating linearized minimization (PALM) algorithm for (7) is also proposed in [7]. Since this algorithm is based on matrix factorization (MF), we name it MF-QRPCA for abbreviation.

B. Problem Formulation

In the Q-TRPCA problem, measurements are in the form of

$$\mathcal{Y}_{ijk} = Q(\widetilde{\mathcal{L}}_{ijk} + \widetilde{\mathcal{C}}_{ijk} + \mathcal{N}_{ijk}) \tag{14}$$

for $\forall(i,j,k) \in [n_1]\times[n_2]\times[n_3]$, where $Q(\cdot)$ is a K-level quantization operator defined by (3), $\widetilde{\boldsymbol{\mathcal{L}}} = (\mathcal{L}_{ijk}) \in \mathbb{R}^{n_1\times n_2\times n_3}$ is the actual tensor whose tubal rank does not exceed r, $\widetilde{\boldsymbol{\mathcal{C}}} = (\widetilde{\mathcal{C}}_{ijk}) \in \mathbb{R}^{n_1\times n_2\times n_3}$ is a sparse tensor that has at most s nonzero entries, and $\boldsymbol{\mathcal{N}} = (\mathcal{N}_{ijk}) \in \mathbb{R}^{n_1\times n_2\times n_3}$ denotes the dither tensor that has i.i.d. entries generated from a cumulative distribution function $g(\cdot)$. For similar reasons to Q-RPCA, we assume $\|\widetilde{\boldsymbol{\mathcal{L}}}\|_\infty \le \alpha$ and $\|\widetilde{\boldsymbol{\mathcal{C}}}\|_\infty \le \alpha$ for some constant $\alpha > 0$ to prevent excessive "spikiness" of the tensors. Similar techniques have been used in [41], [42], [45] when establishing the theoretical guarantees for tensor completion from quantized and partial measurements. Define $\widetilde{\boldsymbol{\mathcal{X}}} = \widetilde{\boldsymbol{\mathcal{L}}} + \widetilde{\boldsymbol{\mathcal{C}}}$ and recall the definition (5). We have that

$$\mathcal{Y}_{ijk} = l \text{ with probability } f_l(\widetilde{\mathcal{X}}_{ijk}),$$

where $\sum_{l=1}^K f_l(\widetilde{\mathcal{X}}_{ijk}) = 1$ for $\forall(i,j,k) \in [n_1]\times[n_2]\times[n_3]$. The problems we consider are now stated as follows.

(P1) Given the quantized observations $\boldsymbol{\mathcal{Y}}$ obtained via (14), noise distribution function $g(\cdot)$, and quantization boundaries $\{\omega_0, \cdots, \omega_K\}$, how well can we recover the true tensor $\widetilde{\boldsymbol{\mathcal{L}}}$ and the sparse component $\widetilde{\boldsymbol{\mathcal{C}}}$?

(P1) is a natural extension of the Q-RPCA problem to tensors. We hope to improve recovery performance by moderately adjusting the measurement model (14) without increasing the quantization level. To this end, we introduce independently repeated dithers in the measurement model (14). More specifically, for each $t \in [W]$, measurement $\mathcal{Y}^t_{ijk}$ is obtained by adding independent dither $\mathcal{N}^t_{ijk}$ to $\widetilde{\mathcal{L}}_{ijk} + \widetilde{\mathcal{C}}_{ijk}$ and then applying quantization using operator $Q(\cdot)$ defined in (3), i.e.,

$$\mathcal{Y}^t_{ijk} = Q(\widetilde{\mathcal{L}}_{ijk} + \widetilde{\mathcal{C}}_{ijk} + \mathcal{N}^t_{ijk}) \tag{15}$$

for $\forall(i,j,k) \in [n_1]\times[n_2]\times[n_3]$ and $\forall t \in [W]$. Recalling the definition (5), for $\forall t \in [W]$ and $\forall(i,j,k) \in [n_1]\times[n_2]\times[n_3]$, we also have

$$\mathcal{Y}^t_{ijk} = l \text{ with probability } f_l(\widetilde{\mathcal{X}}_{ijk}).$$

We note that the measurement model (14) is a special case of (15) when $W = 1$. For observations (15), a more generalized problem of Q-TRPCA is stated as below.

(P2) Given the quantized observations $[\boldsymbol{\mathcal{Y}}^1, \boldsymbol{\mathcal{Y}}^2, \cdots, \boldsymbol{\mathcal{Y}}^W]$ obtained via (15), noise distribution $g(\cdot)$, and quantization function $Q(\cdot)$ with quantization boundaries $\{\omega_0, \cdots, \omega_K\}$, how well can we recover the true tensor $\widetilde{\boldsymbol{\mathcal{L}}}$ and the sparse component $\widetilde{\boldsymbol{\mathcal{C}}}$? How does the recovery error change with the increase in the value of W?

(P2) will be addressed in the next section. Setting $W = 1$ in (P2) trivially implies (P1).

IV. Main Theoretical Results

We now state our main results. First, using a constrained maximum log-likelihood approach, we propose two optimization methods to recover the low-rank component and the sparse component from observations (15). After that, theoretical guarantees for both methods are provided.

A. Optimization Methods

Consider the negative log-likelihood function

$$F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) = -\sum_{i,j,k}\sum_{t=1}^{W}\sum_{l=1}^{K} \mathbb{1}_{\{\mathcal{Y}^t_{ijk}=l\}} \log(f_l(\mathcal{X}_{ijk})), \tag{16}$$

where $\mathscr{Y} := [\boldsymbol{\mathcal{Y}}^1, \cdots, \boldsymbol{\mathcal{Y}}^W]$. Our first claim is that the estimates of $(\widetilde{\boldsymbol{\mathcal{L}}}, \widetilde{\boldsymbol{\mathcal{C}}})$ can be obtained from (15) by solving the following objective

$$(\boldsymbol{\mathcal{L}}^\#, \boldsymbol{\mathcal{C}}^\#) = \arg\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}}) \text{ s.t. } (\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}, \tag{17}$$

where the feasible set $\mathbb{K}_{r,s}$ is defined as

$$\mathbb{K}_{r,s} := \big\{(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) : \boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}} \in \mathbb{R}^{n_1\times n_2\times n_3}, \|\boldsymbol{\mathcal{L}}\|_\infty \le \alpha, \\ \|\boldsymbol{\mathcal{C}}\|_\infty \le \alpha, \mathrm{rank_t}(\boldsymbol{\mathcal{L}}) \le r, \|\boldsymbol{\mathcal{C}}\|_0 \le s\big\}. \tag{18}$$

In Section IV-B, we will analyze the recovery accuracy of the global minimizer of (17). Moreover, we give a lower bound on the best achievable recovery error from unquantized noisy measurements, which does not depend on specific recovery algorithms. We will see that the lower bound is nearly in the same order as the theoretical recovery accuracy obtained by solving (17). This means that the performance of the global minimizer of (17) is nearly order-wise optimal. In this sense, the problem (17) is important for our theoretical analysis. Though the function (16) is convex in $\boldsymbol{\mathcal{X}}$ since $f_l(x)$ is log-concave according to our assumptions, the optimization problem (17) is nonconvex due to the nonconvexity of $\mathbb{K}_{r,s}$. In principle, it can be solved by searching over all critical points whether the objective function reaches the minimum. A critical point is a value in the domain where all partial derivatives are zero. But, clearly this approach is not efficient. For this reason, we next propose a convex programming problem to approximate $(\widetilde{\boldsymbol{\mathcal{L}}},\widetilde{\boldsymbol{\mathcal{C}}})$. Consider any tensors $(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}$. Note that applying Proposition 3 implies that

$$\|\boldsymbol{\mathcal{L}}\|_* = \frac{1}{n_3}\|\bar{\boldsymbol{L}}\|_* \le \frac{1}{n_3}\sqrt{rn_3}\|\bar{\boldsymbol{L}}\|_F = \sqrt{r}\|\boldsymbol{\mathcal{L}}\|_F \\ \le \sqrt{rn_1n_2n_3}\|\boldsymbol{\mathcal{L}}\|_\infty \le \alpha\sqrt{rn_1n_2n_3}. \tag{19}$$

Note also that $\|\boldsymbol{\mathcal{C}}\|_0 \le s$ and $\|\boldsymbol{\mathcal{C}}\|_\infty \le \alpha$ implies $\|\boldsymbol{\mathcal{C}}\|_1 \le \alpha s$. Thus, we have $(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) \in \mathbb{K}^*_{r,s}$, where

$$\mathbb{K}^*_{r,s} := \big\{(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) : \boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}} \in \mathbb{R}^{n_1\times n_2\times n_3}, \|\boldsymbol{\mathcal{L}}\|_\infty \le \alpha, \\ \|\boldsymbol{\mathcal{C}}\|_\infty \le \alpha, \|\boldsymbol{\mathcal{L}}\|_* \le \alpha\sqrt{rn_1n_2n_3}, \|\boldsymbol{\mathcal{C}}\|_1 \le \alpha s\big\}. \tag{20}$$

Considering $\mathbb{K}^*_{r,s}$ as the feasible set, we propose the following optimization objective

$$(\boldsymbol{\mathcal{L}}^*,\boldsymbol{\mathcal{C}}^*) = \arg\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}) \text{ s.t. } (\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) \in \mathbb{K}^*_{r,s}. \tag{21}$$

Problem (21) is respectively convex in $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{C}}$ when the other one is fixed. We also analyze the recovery accuracy of the global minimizer of (21) in Section IV-B. Intuitively, solving (21) will not obtain better theoretical recovery accuracy than solving (17) since a weaker assumption on the low-rankness has been chosen. However, this method is more robust to sparse noises as we show in Section VI.

When $n_3 = 1$ and $W = 1$, our optimization method (17) reduces to (7). Thus (7) is also guaranteed by our theoretical results and (17) can be regarded as a tensor extension of (7). Note that the optimization method (21) does not reduce to (6) when $n_3 = 1$ and $W = 1$ since we add entry-wise infinity norm constraints for a well-posed theoretical analysis. The entry-wise infinity norm constraints usually do not change much in practice [7], [8], [16].

B. Theoretical Recovery Guarantee of the Proposed Methods

With assumptions (8) and (9), we have the following theoretical guarantee on the recovery accuracy of the global minimizer of (17).

Theorem 1. *Suppose that $(\widetilde{\boldsymbol{\mathcal{L}}},\widetilde{\boldsymbol{\mathcal{C}}}) \in \mathbb{K}_{r,s}$, $f_l(x)$ defined by (5) is strictly log-concave in x for any $l \in [K]$, and the observations $\mathscr{Y} = [\boldsymbol{\mathcal{Y}}^1,\cdots,\boldsymbol{\mathcal{Y}}^W]$ are generated as in (15). Then with probability at least $1 - 1/((n_1+n_2)n_3)$, for any global minimizer $(\boldsymbol{\mathcal{L}}^\#,\boldsymbol{\mathcal{C}}^\#)$ of (17), it holds that*

$$\frac{\|\boldsymbol{\mathcal{X}}^\# - \widetilde{\boldsymbol{\mathcal{X}}}\|_F}{\sqrt{n_1n_2n_3}} \le \min\left\{2\alpha\left(1+\sqrt{\frac{2s}{n_1n_2n_3}}\right), U_\alpha\right\}, \tag{22}$$

where $\boldsymbol{\mathcal{X}}^\# = \boldsymbol{\mathcal{L}}^\# + \boldsymbol{\mathcal{C}}^\#$, $\widetilde{\boldsymbol{\mathcal{X}}} = \widetilde{\boldsymbol{\mathcal{L}}} + \widetilde{\boldsymbol{\mathcal{C}}}$, and

$$U_\alpha = \max\left\{\frac{24L_\alpha\sqrt{r\log(d)}}{\gamma_\alpha\sqrt{Wn_{(2)}}}, \right. \\ \left. 4\sqrt{\frac{3\alpha L_\alpha\sqrt{2rsn_{(1)}\log(d)}}{\gamma_\alpha\sqrt{W}n_1n_2n_3} + \frac{\alpha L_\alpha s}{\gamma_\alpha n_1n_2n_3}}\right\}$$

with $d := (n_1+n_2)n_3$. Furthermore, with the same probability, we have

$$\frac{\|\boldsymbol{\mathcal{L}}^\# - \widetilde{\boldsymbol{\mathcal{L}}}\|_F}{\sqrt{n_1n_2n_3}} \\ \le \min\left\{2\alpha\left(1+2\sqrt{\frac{2s}{n_1n_2n_3}}\right), U_\alpha + 2\alpha\sqrt{\frac{2s}{n_1n_2n_3}}\right\}. \tag{23}$$

Theorem 1 establishes the recovery error when the measurements before quantization are sparsely corrupted. As long as s is at most $\Theta(n_{(1)}n_3)$, *i.e.*, the number of corrupted observations per tube of the tensor is bounded, then we have

$$\frac{\|(\boldsymbol{\mathcal{L}}^\# + \boldsymbol{\mathcal{C}}^\#) - (\widetilde{\boldsymbol{\mathcal{L}}} + \widetilde{\boldsymbol{\mathcal{C}}})\|_F}{\sqrt{n_1n_2n_3}} \le \mathcal{O}\left(\sqrt{\frac{r\log(d)}{Wn_{(2)}}}\right), \tag{24}$$

and

$$\frac{\|\boldsymbol{\mathcal{L}}^\# - \widetilde{\boldsymbol{\mathcal{L}}}\|_F}{\sqrt{n_1n_2n_3}} \le \mathcal{O}\left(\sqrt{\frac{r\log(d)}{Wn_{(2)}}}\right). \tag{25}$$

Note that the Frobenius norm of $\boldsymbol{\mathcal{X}}$ is on the same order as $\sqrt{n_1n_2n_3}$. Therefore, by dividing the absolute error by $\sqrt{n_1n_2n_3}$, the left-hand side of (22) is on the same order of the relative error $\|\boldsymbol{\mathcal{X}}^\# - \widetilde{\boldsymbol{\mathcal{X}}}\|_F / \|\widetilde{\boldsymbol{\mathcal{X}}}\|_F$. When the sparse corruption does not exist, i.e., $\boldsymbol{\mathcal{C}} = \boldsymbol{0}$, the optimization objective (17) recovers the underlying tensor from quantized measurements. If s is higher order than $n_{(1)}n_3$, the recovery error will be $\mathcal{O}(\sqrt{s/(n_1n_2n_3)})$. The recovery error still decreases to zero as the dimension increases provided $s \le \mathcal{O}((n_1n_2n_3)^{1-\alpha})$ for some $\alpha > 0$. In addition, the recovery error decreases with the increase of W, which will also be verified by numerical experiments in Section VI-A. The conclusion is meaningful since coarse quantization has high running speed and low cost, which can be executed more times than accurate quantization to compensate for the loss of information in the quantization process. We may also consider how the quantization level K affects the recovery error. Intuitively, the recovery error decreases when the quantization level K increases, which is empirically verified by numerical experiments in Section VI-A. Unfortunately, the error bounds in Theorem 1 do not capture this property explicitly. As is shown in Theorem 1, the recovery error decreases when L_α/γ_α decreases. However,

the dependence of L_α/γ_α on K cannot be characterized in a closed-form expression without giving noise distribution $g(x)$ and quantization boundaries ω_l's. When the noise follows the logistic distribution $\Phi_{log}(\frac{x}{\sigma}) = 1/(1+e^{-x/\sigma})$, L_α/γ_α is minimized when $K = 2$ [7], which contradicts the experimental results in Section VI-A. This counter-intuitive result is due to the artifacts in the proof, which are built upon sufficient but not necessary conditions to study the worst-case performance.

To illustrate the connection between our work and Q-RPCA, we can compare Theorem 1 with results in [7]. When $n_3 = 1$ and $W = 1$, Theorem 1 reduces to a theoretical result for Q-RPCA. Unfortunately, our results are slightly inferior to that of [7], and they differ by a factor $\sqrt{\log(n_1+n_2)}$ though it is far less than $\sqrt{n_{(2)}}$. This is because the tool that [7] used in their proof to bound the matrix spectral norm cannot be directly extended in the framework of t-SVD. To conquer this issue, we use other tools when we bound the tensor spectral norm.

In order to evaluate the quality of the upper bound on the recovery error derived in Theorem 1, we will use information-theoretic methods to obtain a lower bound on the recovery error of any method for the set $\mathbb{K}_{r,s}$ even when the measurements are not quantized. As we will see in Theorem 2, the lower bound nearly matches the upper bound on the convergence rate, up to a logarithmic factor in $(n_1 + n_2)n_3$. Before introducing Theorem 2, we require a few assumptions. First, we assume that the random noises are generated from a Gaussian distribution with zero mean and variances σ^2. Second, we suppose that s is at most $\Theta(n_1n_2n_3)$, and n_1, n_2, and n_3 are sufficiently large. To simplify the presentation, we quantify this region as

$$\frac{s}{n_1n_2n_3} \leq C_0 \text{ and } \frac{r}{n_{(2)}}(n_1n_2n_3 - s) \geq 64, \tag{26}$$

where $C_0 > 0$ is a constant. These assumptions are mild and do not contradict the settings of Theorem 1.

Theorem 2. *Suppose $\boldsymbol{\mathcal{N}}^t \in \mathbb{R}^{n_1\times n_2\times n_3}$ for $t \in [W]$ are noise tensors containing i.i.d. entries generated from $N(0, \sigma^2)$, and* (26) *holds. Let $\boldsymbol{\mathcal{Y}}^t$ be obtained via $\boldsymbol{\mathcal{Y}}^t = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{N}}^t$ for all $t \in [W]$. Consider any algorithm that, for any $(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}$, takes $\mathscr{Y} = [\boldsymbol{\mathcal{Y}}^1, \cdots, \boldsymbol{\mathcal{Y}}^W]$ as the input and returns $(\hat{\boldsymbol{\mathcal{L}}}, \hat{\boldsymbol{\mathcal{C}}})$ as an estimate of $(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}})$. Then there must exist $(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}$ such that with probability at least $3/4$, it holds that*

$$\frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_F}{\sqrt{n_1n_2n_3}} \geq \min\left\{C_2, \frac{C_1\sigma}{\sqrt{W}}\sqrt{\frac{rn_{(1)}n_3 - r\lfloor\frac{s}{n_{(2)}}\rfloor - 64}{n_1n_2n_3 - n_{(2)}\lfloor\frac{s}{n_{(2)}}\rfloor}}\right\} \tag{27}$$

for some fixed constants C_1 and C_2, where $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}}$, $\hat{\boldsymbol{\mathcal{X}}} = \hat{\boldsymbol{\mathcal{L}}} + \hat{\boldsymbol{\mathcal{C}}}$, $C_1 < \sqrt{\frac{1-2C_0}{256}}$, and $C_2 = \alpha\sqrt{\frac{(1-2C_0)}{8}}$.

Note that if s is no more than $\Theta(n_1n_2n_3)$, then (27) implies

$$\frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_F}{\sqrt{n_1n_2n_3}} \geq \Theta\left(\sqrt{\frac{r}{n_{(2)}W}}\right). \tag{28}$$

This means that for the set $\mathbb{K}_{r,s}$, even without quantization, the worst-case minimum possible recovery error from noisy measurements $\boldsymbol{\mathcal{Y}}^t = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{N}}^t$ is at least on the order of $\Theta\left(\sqrt{r/(n_{(2)}W)}\right)$, provided s is at most $\Theta(n_1n_2n_3)$. From (24), the recovery error by solving the global minimizer of (17) is at most $\Theta\left(\sqrt{r\log(d)/(n_{(2)}W)}\right)$ with $d = (n_1+n_2)n_3$, as long as s is at most $\Theta(n_{(1)}n_3)$. Theorem 2 is algorithm-independent. Therefore, we conclude that the recovery error of solving (17) is nearly order-optimal in the sense that it is almost on the same order as the minimum possible error obtained when the recovery is from unquantized measurements, up to a logarithm factor in $(n_1 + n_2)n_3$, provided s is at most $\Theta(n_{(1)}n_3)$. This difference may come from the artifacts in our proof techniques.

Theorem 1 combined with Theorem 2 describes the recovery performance of solving the optimization problem (17). We also want to know how the recovery error changes when we consider problem (21), which will be addressed in Theorem 3. Before providing the upper bound, we need define a positive constant K_α such that

$$K_\alpha \leq \inf_{\substack{u,v\in\mathbb{R}\\ |u|\leq 2\alpha\\ |v|\leq 2\alpha}} \left(\sum_{l=1}^{K}\left(\sqrt{f_l(u)} - \sqrt{f_l(v)}\right)^2/|u-v|^2\right), \tag{29}$$

where f_l is given by (5). The upper bound of (29) relies on α, g, and ω_i's ($i \in \{0, 1, \cdots, K\}$) but is independent of n_1, n_2 and n_3. The same assumption has also been used in [14].

Theorem 3. *Assume that $(\widetilde{\boldsymbol{\mathcal{L}}}, \widetilde{\boldsymbol{\mathcal{C}}}) \in \mathbb{K}^*_{r,s}$, $\widetilde{\boldsymbol{\mathcal{X}}} = \widetilde{\boldsymbol{\mathcal{L}}} + \widetilde{\boldsymbol{\mathcal{C}}}$, $f_l(x)$ defined by* (5) *is strictly log-concave in x for any $l \in [K]$, and the observations $\mathscr{Y} = [\boldsymbol{\mathcal{Y}}^1, \cdots, \boldsymbol{\mathcal{Y}}^W]$ are generated as in* (15)*. Let L_α and K_α be defined as in* (9) *and* (29)*, respectively. Let $(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{C}}^*)$ be the solution to* (21) *and $\boldsymbol{\mathcal{X}}^* = \boldsymbol{\mathcal{L}}^* + \boldsymbol{\mathcal{C}}^*$. Then with probability at least $1 - \frac{1}{(n_1+n_2)n_3}$, we have*

$$\frac{\|\boldsymbol{\mathcal{X}}^* - \widetilde{\boldsymbol{\mathcal{X}}}\|_F^2}{n_1n_2n_3} \leq \frac{2C_0\alpha L_\alpha}{K_\alpha}\left(\sqrt{\frac{r\log^2(d)}{n_{(2)}W}} + \frac{s\sqrt{\log(d)}}{n_{(2)}\sqrt{Wn_3n_{(1)}}}\right), \tag{30}$$

where C_0 is an absolute constant. Furthermore, with the same probability, we have

$$\frac{\|\boldsymbol{\mathcal{L}}^* - \widetilde{\boldsymbol{\mathcal{L}}}\|_F}{\sqrt{n_1n_2n_3}} \leq 2\alpha\sqrt{\frac{2s}{n_1n_2n_3}} + \sqrt{2C_0\alpha\frac{L_\alpha}{K_\alpha}\left(\sqrt{\frac{r\log^2(d)}{n_{(2)}W}} + \frac{s\sqrt{\log(d)}}{n_{(2)}\sqrt{Wn_3n_{(1)}}}\right)}.$$

From Theorem 3, we have that

$$\frac{\|(\boldsymbol{\mathcal{L}}^* + \boldsymbol{\mathcal{C}}^*) - (\widetilde{\boldsymbol{\mathcal{L}}} + \widetilde{\boldsymbol{\mathcal{C}}})\|_F}{\sqrt{n_1n_2n_3}} \leq \mathcal{O}\left(\left(\frac{r\log^2(d)}{Wn_{(2)}}\right)^{1/4}\right), \tag{31}$$

and

$$\frac{\|\boldsymbol{\mathcal{L}}^* - \widetilde{\boldsymbol{\mathcal{L}}}\|_F}{\sqrt{n_1n_2n_3}} \leq \mathcal{O}\left(\left(\frac{r\log^2(d)}{Wn_{(2)}}\right)^{1/4}\right), \tag{32}$$

as long as s is at most $\Theta(\sqrt{n_1n_2n_3})$. If $\Theta(\sqrt{n_1n_2n_3}) \leq s \leq \Theta(n_{(2)}\sqrt{n_{(1)}n_3})$, both (31) and (32) are upper bounder by $\mathcal{O}((\frac{s^2\log(d)}{Wn_{(1)}n_3n_{(2)}^2})^{1/4})$.

Remark 1. *Because $f_l(x)$ is assumed to be differentiable, the Mean Value Theorem indicates that the right-hand side of* (29) *is more than*

$$\inf_{|\xi|\leq 2\alpha}\left(\sum_{l=1}^{K}\frac{(\dot{f}_l(\xi))^2}{4f_l(\xi)}\right)$$

where $\dot{f}_l(x)$ denotes the first order derivative of f_l with respect to x. Recalling $f_l(x)\leq 1$, we can define K_α as

$$K_\alpha := \frac{1}{4}\sum_{l=1}^{K}\left(\inf_{|\xi|\leq 2\alpha}(\dot{f}_l(\xi))^2\right) \tag{33}$$

for simplicity. The assumption (33) *is well-defined for many distribution functions such as the logistic distribution $g(x) := \Phi_{log}(\frac{x}{\sigma}) = 1/(1+e^{-x/\sigma})$ and the Gaussian distribution $g(x) := \Phi_{norm}(\frac{x}{\sigma})$.*

Next, for the set $\mathbb{K}^*_{r,s}$, we also obtain the lower bound on the recovery error of any method when the measurements are not quantized. Without loss of generality, we assume that $\alpha \geq 1$ and

$$\alpha^2(rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor) \geq C_0 \tag{34}$$

holds for some constant $C_0 > 0$.

Theorem 4. *Suppose $\boldsymbol{\mathcal{N}}^t \in \mathbb{R}^{n_1\times n_2\times n_3}$ for $t\in[W]$ are noise tensors containing i.i.d. entries generated from $N(0,\sigma^2)$, and* (34) *holds. Let $\boldsymbol{\mathcal{Y}}^t$ be obtained via $\boldsymbol{\mathcal{Y}}^t = \boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{N}}^t$ for all $t\in[W]$. Consider any algorithm that, for any $(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}})\in\mathbb{K}^*_{r,s}$, takes $\mathscr{Y} = [\boldsymbol{\mathcal{Y}}^1,\cdots,\boldsymbol{\mathcal{Y}}^W]$ as the input and returns $(\hat{\boldsymbol{\mathcal{L}}},\hat{\boldsymbol{\mathcal{C}}})$ as an estimate of $(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}})$. Then there must exist $(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}})\in\mathbb{K}^*_{r,s}$ such that with probability at least $3/4$, it holds that*

$$\frac{\|\hat{\boldsymbol{\mathcal{X}}}-\boldsymbol{\mathcal{X}}\|_F^2}{n_1n_2n_3} \geq \min\left\{C_1, C_2\alpha\sigma\sqrt{\frac{rn_{(1)}n_3 - r\lfloor\frac{s}{n_{(2)}}\rfloor}{W(n_1n_2n_3 - n_{(2)}\lfloor\frac{s}{n_{(2)}}\rfloor)}}\right\} \tag{35}$$

for some fixed constants C_1 and C_2, where $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}$, $\hat{\boldsymbol{\mathcal{X}}} = \hat{\boldsymbol{\mathcal{L}}}+\hat{\boldsymbol{\mathcal{C}}}$ as long as the right-hand slide of (35) *exceeds $r\alpha^2/n_{(2)}$.*

The requirement that the right-hand slide of (35) exceeds $r\alpha^2/n_{(2)}$ is satisfied provided

$$r \leq C_3\frac{\min\{1,\sigma^2\}\min\{n_1,n_2\}}{\alpha^2}$$

holds for a fixed constant C_3. Note that if s is no more than $\Theta(n_1n_2n_3)$, then (35) implies

$$\frac{\|\hat{\boldsymbol{\mathcal{X}}}-\boldsymbol{\mathcal{X}}\|_F}{\sqrt{n_1n_2n_3}} \geq \Theta\left(\left(\frac{r}{n_{(2)}W}\right)^{\frac{1}{4}}\right). \tag{36}$$

Thus, for the relaxed set $\mathbb{K}^*_{r,s}$, the worst-case minimum possible recovery error from noisy measurements $\boldsymbol{\mathcal{Y}}^t = \boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{N}}^t$ is at least on the order of $\Theta\left(\sqrt[4]{r/(n_{(2)}W)}\right)$, provided s is at most $\Theta(n_1n_2n_3)$. From (30), the recovery error by solving the global minimizer of (21) is at most $\Theta\left(\sqrt[4]{r\log(d)^2/(n_{(2)}W)}\right)$ with $d=(n_1+n_2)n_3$, as long as s is at most $\Theta(n_1n_2n_3)$. Therefore, for the set $\mathbb{K}^*_{r,s}$, we conclude that the recovery error of solving (21) is nearly order-optimal in the sense that it is almost on the same order as the minimum possible error given by (36), up to the half-power of $\log(d)$, provided s is at most $\Theta(n_1n_2n_3)$.

The theoretical upper bound on the recovery error in Theorem 3 decreases more slowly than that in Theorem 1, which is reasonable since we make a weaker assumption on the low-rankness of tensors. Moreover, we will see in Section VI-C that solving (21) usually obtains more robust recovery performance when the corruption rate is low, as compared with solving (17). In addition, solving (21) often performs better in practical applications, as we will see in Section VII.

V. Q-TRPCA Algorithms

We have proposed problems (17) and (21) in Section IV-A and provided theoretical guarantees on the estimation errors of the global minimizers of (17) and (21) in Section IV-B. In this section, we consider algorithms to solve (17) and (21), respectively.

Since the algorithms proposed in this section are applications of the PALM method [53], we first review PALM to make this paper self-contained. PALM aims to solve a broad class of nonconvex-nonsmooth minimization problems of the form

$$\min_{\boldsymbol{x},\boldsymbol{y}}\Psi(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x})+g(\boldsymbol{y})+H(\boldsymbol{x},\boldsymbol{y}), \tag{37}$$

over all $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^n\times\mathbb{R}^m$, where the functions $f:\mathbb{R}^n\to(-\infty,\infty]$ and $g:\mathbb{R}^m\to(-\infty,\infty]$ are proper and lower semicontinuous functions and H is a smooth function. PALM can be seen as alternating the steps of the proximal forward-backward (PFB) scheme to alternately update $\boldsymbol{x}$ and $\boldsymbol{y}$. It is well-known that the PFB scheme for minimizing the sum of a smooth function h and a nonsmooth function σ can simply be viewed as the proximal regularization of h linearized at a given point $\boldsymbol{v}^k$ [53], i.e.,

$$\boldsymbol{v}^{k+1}\in\arg\min_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\langle\boldsymbol{v}-\boldsymbol{v}^k,\nabla h(\boldsymbol{v}^k)\rangle+\frac{\|\boldsymbol{v}-\boldsymbol{v}^k\|_2^2}{2\tau}+\sigma(\boldsymbol{v})\right\}. \tag{38}$$

Given $\boldsymbol{x}\in\mathbb{R}^n$ and $\tau>0$, the proximal map associated to the map σ is defined by

$$\mathrm{prox}_\tau^\sigma(\boldsymbol{v}) := \arg\min\left\{\sigma(\boldsymbol{u})+\frac{1}{2\tau}\|\boldsymbol{u}-\boldsymbol{v}\|_2^2 : \boldsymbol{u}\in\mathbb{R}^n\right\}, \tag{39}$$

that is, combining (38) with (39), one has

$$\boldsymbol{v}^{k+1}\in\mathrm{prox}_\tau^\sigma\left(\boldsymbol{v}^k-\tau\nabla h(\boldsymbol{v}^k)\right).$$

Adopting this scheme on (37) obtains the main updates of PALM. For each iteration $t = 0,1,\cdots$, PALM alternates updating two blocks $(\boldsymbol{x},\boldsymbol{y})$ by the following steps: (i) fix $\boldsymbol{y}^t$ and compute $\boldsymbol{x}^{t+1}$ via $\boldsymbol{x}^{k+1}\in\mathrm{prox}^f_{\tau_{\boldsymbol{x}}}(\boldsymbol{x}^t-\tau_{\boldsymbol{x}}\nabla_{\boldsymbol{x}}H(\boldsymbol{x}^t,\boldsymbol{y}^t))$; (ii) fix $\boldsymbol{x}^{t+1}$ and compute $\boldsymbol{y}^{t+1}$ via $\boldsymbol{y}^{k+1}\in\mathrm{prox}^g_{\tau_{\boldsymbol{y}}}\left(\boldsymbol{x}^t-\tau_{\boldsymbol{y}}\nabla_{\boldsymbol{y}}H(\boldsymbol{x}^{t+1},\boldsymbol{y}^t)\right)$. Here $\tau_{\boldsymbol{x}}$ and $\tau_{\boldsymbol{y}}$ are step sizes which are usually selected via a backtracking line search using Armijo's rule. Taking $\tau_{\boldsymbol{x}}$ as an example, we first fix a parameter $\beta\in(0,1)$ and start with $\tau_{\boldsymbol{x}}=1$. Then we continuously update $\tau_{\boldsymbol{x}}$ by $\tau_{\boldsymbol{x}}=\beta\tau_{\boldsymbol{x}}$

until it holds that $H(\boldsymbol{x}^t - \tau_{\boldsymbol{x}}\nabla_{\boldsymbol{x}}H(\boldsymbol{x}^t,\boldsymbol{y}^t),\boldsymbol{y}^t) \leq H(\boldsymbol{x}^t) - \tau_{\boldsymbol{x}}\|\nabla_{\boldsymbol{x}}H(\boldsymbol{x}^t,\boldsymbol{y}^t)\|_F^2/2$. PALM has been proven to converge to critical points of problem (37) in [53]. A critical point is a value in the domain where all partial derivatives are zero. Note that the choice of two blocks of variables in (37) is for the sake of simplicity of exposition. Indeed, the results hold true for a finite number of block-variables [53].

A. Tensor Factorization based Quantized Robust Principal Component Analysis

We next consider (17). For any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ in the set $\mathbb{K}_f$ where $\mathbb{K}_f := \{\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} : (\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}\}$, Proposition 2 implies that $\boldsymbol{\mathcal{X}}$ can been written as

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}},$$

where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1\times r\times n_3}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{r\times n_2\times n_3}$, and $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ satisfying $\|\boldsymbol{\mathcal{C}}\|_0 \leq s$. Therefore, (17) can be recast as

$$(\hat{\boldsymbol{\mathcal{U}}},\hat{\boldsymbol{\mathcal{V}}},\hat{\boldsymbol{\mathcal{C}}}) = \underset{\boldsymbol{\mathcal{U}},\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{C}}}{\arg\min} F_{\mathscr{Y}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) + K_1(\boldsymbol{\mathcal{C}}), \tag{40}$$

where $K_1(\cdot)$ is defined as

$$K_1(\boldsymbol{\mathcal{C}}) = \begin{cases} 0, \text{ if } \|\boldsymbol{\mathcal{C}}\|_0 \leq s, \\ +\infty, \text{ if } \|\boldsymbol{\mathcal{C}}\|_0 > s, \end{cases} \tag{41}$$

to guarantee the sparsity of $\boldsymbol{\mathcal{C}}$. Thus $K_1(\boldsymbol{\mathcal{C}})$ is an indicator function of a semi-algebraic set. We discard the constraints $\|\boldsymbol{\mathcal{L}}\|_\infty \leq \alpha$ and $\|\boldsymbol{\mathcal{C}}\|_\infty \leq \alpha$ to simplify the algorithm. We note that ignoring these constraints is rather innocuous in practical cases since the numerical results are not affected much, despite the key role they played in our theoretical analysis. A similar treatment has also been used in [7] and [8]. In fact, we can assume that α is large enough so that no entry can exceed it, and thus the infinity norm based constraints in (17) are not necessary for our algorithm.

Our algorithm borrows ideas from [7], which is based on the following properties of the matrix gradient. Consider matrices $\boldsymbol{U} \in \mathbb{R}^{m\times\ell}$, $\boldsymbol{V} \in \mathbb{R}^{\ell\times n}$, $\boldsymbol{C} \in \mathbb{R}^{m\times n}$, and a map $\mu(\boldsymbol{X}) : \mathbb{R}^{m\times n} \to \mathbb{R}$ that applies to $\boldsymbol{X}$ in an element-wise manner, i.e., $\mu(\boldsymbol{\mathcal{X}}) = \sum_{i,j}\psi(X_{ij})$ with $\psi : \mathbb{R} \to \mathbb{R}$ a given map. Then,

$$\begin{aligned} \nabla_{\boldsymbol{U}}\mu(\boldsymbol{UV}+\boldsymbol{C}) &= \nabla_{\boldsymbol{X}}\mu(\boldsymbol{X})\boldsymbol{V}^T|_{\boldsymbol{X}=\boldsymbol{UV}+\boldsymbol{C}}, \\ \nabla_{\boldsymbol{V}}\mu(\boldsymbol{UV}+\boldsymbol{C}) &= \boldsymbol{U}^T\nabla_{\boldsymbol{X}}\mu(\boldsymbol{X})|_{\boldsymbol{X}=\boldsymbol{UV}+\boldsymbol{C}}, \\ \nabla_{\boldsymbol{C}}\mu(\boldsymbol{UV}+\boldsymbol{C}) &= \nabla_{\boldsymbol{X}}\mu(\boldsymbol{X})|_{\boldsymbol{X}=\boldsymbol{UV}+\boldsymbol{C}}, \end{aligned} \tag{42}$$

and

$$[\nabla_{\boldsymbol{X}}\mu(\boldsymbol{X})]_{ij} = \dot{\psi}(X_{ij}), \tag{43}$$

where $\nabla_{\boldsymbol{M}}\mu(\boldsymbol{M})|_{\boldsymbol{M}=\boldsymbol{T}}$ denotes the value of $\nabla_{\boldsymbol{M}}\mu(\boldsymbol{M})$ when $\boldsymbol{M}=\boldsymbol{T}$. To generalize the idea of [7] to the tensor case, we have to extend the above properties to tensors under the framework of t-SVD. Suppose $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1\times r\times n_3}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{r\times n_2\times n_3}$, and $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{n_1\times n_2\times n_3}$. Let $\mu : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}$ be a map that applies to tensors in an element-wise manner, i.e., $\mu(\boldsymbol{\mathcal{X}}) = \sum_{i,j,k}\psi(\mathcal{X}_{ijk})$ with $\psi : \mathbb{R} \to \mathbb{R}$ being a given map. Thus, with a slight abuse of notation, we can regard μ as a map from $\mathbb{R}^{n_1n_3\times n_2}$ to $\mathbb{R}$ and rewrite $\mu(\boldsymbol{\mathcal{X}})$ as $\mu(\text{unfold}(\boldsymbol{\mathcal{X}}))$.

Note that $\text{unfold}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}) = \text{bcirc}(\boldsymbol{\mathcal{U}})\text{unfold}(\boldsymbol{\mathcal{V}})$. Thus we have

$$\begin{aligned} &\text{unfold}(\nabla_{\boldsymbol{\mathcal{V}}}\mu(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}})) \\ &= \nabla_{\text{unfold}(\boldsymbol{\mathcal{V}})}\mu(\text{bcirc}(\boldsymbol{\mathcal{U}})\text{unfold}(\boldsymbol{\mathcal{V}}) + \text{unfold}(\boldsymbol{\mathcal{C}})) \\ &= \text{bcirc}(\boldsymbol{\mathcal{U}})^T\nabla_{\boldsymbol{X}}\mu(\boldsymbol{X})|_{\boldsymbol{X}=\text{unfold}(\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}})+\text{unfold}(\boldsymbol{\mathcal{C}})} \\ &= \text{bcirc}(\boldsymbol{\mathcal{U}})^T\text{unfold}(\nabla_{\boldsymbol{\mathcal{X}}}\mu(\boldsymbol{\mathcal{X}})|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}), \end{aligned}$$

which implies that

$$\nabla_{\boldsymbol{\mathcal{V}}}\mu(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \boldsymbol{\mathcal{U}}^T * \nabla_{\boldsymbol{\mathcal{X}}}\mu(\boldsymbol{\mathcal{X}})|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}. \tag{44}$$

Taking the tensor transpose and following similar steps, we obtain

$$\nabla_{\boldsymbol{\mathcal{U}}}\mu(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \nabla_{\boldsymbol{\mathcal{X}}}\mu(\boldsymbol{\mathcal{X}}) * \boldsymbol{\mathcal{V}}^T|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}. \tag{45}$$

In addition, it is obvious that

$$\nabla_{\boldsymbol{\mathcal{C}}}\mu(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \nabla_{\boldsymbol{\mathcal{X}}}\mu(\boldsymbol{\mathcal{X}})|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}. \tag{46}$$

We can check that $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$ is in the form of $\sum_{i,j,k}\psi(\mathcal{X}_{ijk})$ where $\psi(\mathcal{X}_{ijk}) := -\sum_{t=1}^{W}\sum_{l=1}^{S}\mathbb{1}_{\{\mathcal{Y}^t_{ijk}=l\}}\log(f_l(\mathcal{X}_{ijk}))$ with f_l given by (5). Therefore, setting $\mu := F_{\mathscr{Y}}$ in (44), (45), and (46) implies

$$\nabla_{\boldsymbol{\mathcal{V}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \boldsymbol{\mathcal{U}}^T * \nabla_{\boldsymbol{\mathcal{X}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}, \tag{47}$$

$$\nabla_{\boldsymbol{\mathcal{U}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \nabla_{\boldsymbol{\mathcal{X}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) * \boldsymbol{\mathcal{V}}^T|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}, \tag{48}$$

and

$$\nabla_{\boldsymbol{\mathcal{C}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}} + \boldsymbol{\mathcal{C}}) = \nabla_{\boldsymbol{\mathcal{X}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})|_{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{U}}*\boldsymbol{\mathcal{V}}+\boldsymbol{\mathcal{C}}}, \tag{49}$$

where

$$[\nabla_{\boldsymbol{\mathcal{X}}}F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]_{ijk} = -\sum_{t=1}^{W}\frac{\dot{f}_{\mathcal{Y}^t_{ijk}}(\mathcal{X}_{ijk})}{f_{\mathcal{Y}^t_{ijk}}(\mathcal{X}_{ijk})}$$

with f_l given by (5) and $\dot{f}_l$ denoting the first order derivative of f_l as defined in (9). $K_1(\boldsymbol{\mathcal{C}})$ in (40) can be achieved by maintaining the largest s entries in the magnitude and setting others to be zero according to Proposition 4 in [53]. Using the framework of PALM, we propose Algorithm 1 to solve (40). Since Algorithm 1 is based on tensor factorization, we name it TF-QRPCA for simplicity. When $n_3 = 1$, TF-QRPCA reduces to MF-QRPCA in [7].

Algorithm 1 highly depends on the estimation of the tubal rank r. We leverage the Akaike Information Criterion (AIC) to choose the tubal rank that minimizes AIC. More specifically, we select r by solving the following problem

$$\hat{r} = \underset{r\in\mathbb{N}_+}{\arg\min}\,\text{AIC}(r) = \underset{r\in\mathbb{N}_+}{\arg\min}\{2F_{\mathscr{Y}}(\boldsymbol{\mathcal{L}}(r)+\boldsymbol{\mathcal{C}}(r)) + 2p\}, \tag{50}$$

where $\mathbb{N}_+$ denotes the set of positive integers, $\boldsymbol{\mathcal{L}}(r)$, $\boldsymbol{\mathcal{C}}(r)$ are the estimates given the tubal rank r, and $p := r(n_1+n_2-r)n_3$ denotes the degree of freedoms of an $n_1\times n_2\times n_3$ tensor with tubal rank r. The choice of AIC aims to balance between the goodness-of-fit for the data and the degree of freedoms in the optimization problem.

Algorithm 1 : TF-QRPCA

Input: Quantized tensor $\mathcal{Y}$, parameters tol, s, r.
Output: $\hat{\mathcal{L}}$, $\hat{\mathcal{C}}$.

1: Initiate: $\mathcal{U}^0 \in \mathbb{R}^{n_1\times r\times n_3}$, $\mathcal{V}^0 \in \mathbb{R}^{r\times n_2\times n_3}$, and zero tensor $\mathcal{C}^0 = \mathbf{0}$.
2: **for** $t = 0, 1, 2, \cdots$, until $\|\mathcal{U}^{t+1}-\mathcal{U}^t\|_F+\|\mathcal{V}^{t+1}-\mathcal{V}^t\|_F+\|\mathcal{C}^{t+1}-\mathcal{C}^t\|_F < tol$ or $t = T$ **do**
3: $\mathcal{U}^{t+1} = \mathcal{U}^t - \tau_{\mathcal{U}}\nabla_{\mathcal{U}}F_{\mathscr{Y}}(\mathcal{U}^t * \mathcal{V}^t + \mathcal{C}^t)$ where $\nabla_{\mathcal{U}}F_{\mathscr{Y}}(\mathcal{U}^t * \mathcal{V}^t + \mathcal{C}^t)$ is computed via (47).
4: $\mathcal{V}^{t+1} = \mathcal{V}^t - \tau_{\mathcal{V}}\nabla_{\mathcal{V}}F_{\mathscr{Y}}(\mathcal{U}^{t+1} * \mathcal{V}^t + \mathcal{C}^t)$ where $\nabla_{\mathcal{V}}F_{\mathscr{Y}}(\mathcal{U}^{t+1} * \mathcal{V}^t + \mathcal{C}^t)$ is computed via (48).
5: $\mathcal{C}^{t+1} = \mathcal{C}^t - \tau_{\mathcal{C}}\nabla_{\mathcal{C}}F_{\mathscr{Y}}(\mathcal{U}^{t+1} * \mathcal{V}^{t+1} + \mathcal{C}^t)$ where $\nabla_{\mathcal{C}}F_{\mathscr{Y}}(\mathcal{U}^{t+1} * \mathcal{V}^{t+1} + \mathcal{C}^t)$ is computed via (49).
6: **if** $\|\mathcal{C}\|_0 > s$ **then**
7: $\mathcal{C}^{t+1}$ only keeps the s largest entries in the absolute magnitude, setting other nonzero entries to be zero.
8: **end if**
9: **end for**
10: **return** $\hat{\mathcal{L}} = \mathcal{U}^{t+1} * \mathcal{V}^{t+1}$ and $\hat{\mathcal{C}} = \mathcal{C}^{t+1}$.

B. Tensor Nuclear Norm Based Quantized Robust Principal Component Analysis

We now propose an algorithm to solve the optimization problem (21). Similar to the above discussion, we remove the infinity-norm constraints to simplify the algorithm and consider the following problem

$$(\hat{\mathcal{L}}, \hat{\mathcal{C}}) = \arg\min_{\mathcal{L},\mathcal{C}} \; F_{\mathscr{Y}}(\mathcal{L}+\mathcal{C}) + N(\mathcal{L}) + K_2(\mathcal{C}), \tag{51}$$

where $N(\mathcal{L})$ is defined as

$$N(\mathcal{L}) = \begin{cases} 0, \text{ if } \|\mathcal{L}\|_* \le \lambda, \\ +\infty, \text{ if } \|\mathcal{L}\|_* > \lambda \end{cases} \tag{52}$$

to promote the low-rankness of $\mathcal{L}$ and $K_2(\mathcal{C})$ is defined as

$$K_2(\mathcal{C}) = \begin{cases} 0, \text{ if } \|\mathcal{C}\|_1 \le \mu, \\ +\infty, \text{ if } \|\mathcal{C}\|_1 > \mu \end{cases} \tag{53}$$

to promote the sparsity of $\mathcal{C}$. Here we do not use the constraints $\|\mathcal{L}\|_* \le \alpha\sqrt{rn_1n_2n_3}$ and $\|\mathcal{C}\|_1 \le \alpha s$ in (52) and (53) since they are too large in practice though they are proper for our theoretical analysis. In fact, only a few entries of the tensor have large absolute values while most entries have relatively small absolute values in many practical cases. Therefore, the parameters λ and μ are usually selected via cross-validation or using prior information of the dataset. Note that $F_{\mathscr{Y}}$ is a smooth function and both $N(\cdot)$ and $K_2(\cdot)$ are proper and lower semicontinuous functions. Therefore, we can use PALM to solve (51). $N(\mathcal{L})$ in (51) can be achieved by computing $\bar{\boldsymbol{L}} = \mathrm{bdiag}(\bar{\mathcal{L}})$, projecting $\bar{\boldsymbol{L}}$ to

$$\{\boldsymbol{Z} \in \mathbb{R}^{n_1n_3\times n_2n_3} | \|\boldsymbol{Z}\|_* \le n_3\lambda\}, \tag{54}$$

and then transforming the resulting block diagonal matrix back to a tensor followed by an inverse DFT. The orthogonal projection onto a matrix nuclear norm ball amounts to singular value soft thresholding. In fact, let $\mathbb{D} := \{\boldsymbol{Z} | \|\boldsymbol{Z}\|_* \le \eta\}$ and $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the SVD of $\boldsymbol{T}$ with $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$. Then the orthogonal projection of $\boldsymbol{T}$ to the set $\mathbb{D}$ is

$$\mathcal{P}_{\mathbb{D}}(\boldsymbol{T}) := \boldsymbol{U}\max\{\boldsymbol{\Sigma} - \xi\boldsymbol{I}, 0\}\boldsymbol{V}^*, \tag{55}$$

where the maximum is taken entrywise with $\xi > 0$ being the smallest value such that $\sum_{i=1}^{p}\max\{\sigma_i - \xi, 0\} \le \eta$. $K_2(\mathcal{C})$ in (51) can be achieved using the algorithm detailed in [54]. Our algorithm performs two steps in iteration: (i) hold the sparse component $\mathcal{C}$ fixed and optimizes the low-rank component $\mathcal{L}$ using PFB; (ii) hold the low-rank part $\mathcal{L}$ fixed and optimizes the sparse component $\mathcal{C}$ using PFB. The procedure is summarized in Algorithm 2 where Lines $3, 4$ corresponds to (i) and Lines $5, 6$ corresponds to (ii). We initiate Algorithm 2 via

$$\mathcal{L}^0_{ijk} = \begin{cases} \dfrac{\omega_l + \omega_{l-1}}{2}, & \text{if } 1 < \mathcal{Y}^1_{ijk} = l < K, \\ \dfrac{\alpha + \omega_{K-1}}{2}, & \text{if } \mathcal{Y}^1_{ijk} = K, \\ \dfrac{-\alpha + \omega_1}{2}, & \text{if } \mathcal{Y}^1_{ijk} = 1. \end{cases} \tag{56}$$

Since Algorithm 2 uses tensor nuclear norm (TNN) to promote the low-rankness, we name it TNN-QRPCA to distinguish it from TF-QRPCA. When $n_3 = 1$, TNN-QRPCA reduces to an algorithm for Q-RPCA. We did not use a BCD based method since, for each outer iteration, it needs to perform an inner iterative procedure when optimizing each subproblem, which increases the computational cost.

Algorithm 2 :TNN-QRPCA

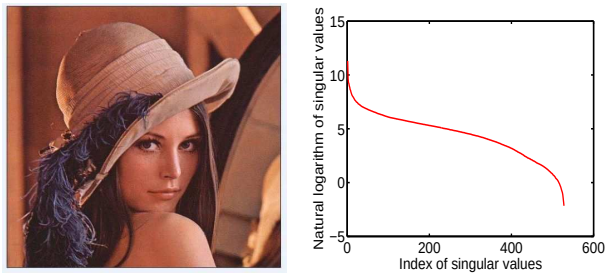
Input: Quantized tensor $\mathcal{Y}$, parameters tol, s, r.
Output: $\hat{\mathcal{L}}$, $\hat{\mathcal{C}}$.

1: Initiate: $\mathcal{L}^0$ via (56), $\mathcal{C}^0 = \mathbf{0}$.
2: **for** $t = 0, 1, 2, \cdots$, until $\|\mathcal{L}^{t+1}-\mathcal{L}^t\|_F+\|\mathcal{C}^{t+1}-\mathcal{C}^t\|_F < tol$ or $t = T$ **do**
3: $\mathcal{L}^{t+1} = \mathcal{L}^t - \tau_{\mathcal{L}}\nabla_{\mathcal{L}}F(\mathcal{L}^t + \mathcal{C}^t)$.
4: Projecting $\mathcal{L}^{t+1}$ to the set $\{\mathcal{T} \in \mathbb{R}^{n_1\times n_2\times n_3} | \|\mathcal{T}\|_* \le \lambda\}$ via (54) and (55).
5: $\mathcal{C}^{t+1} = \mathcal{C}^t - \tau_{\mathcal{C}}\nabla_{\mathcal{C}}F(\mathcal{L}^{t+1} + \mathcal{C}^t)$.
6: Projecting $\mathcal{C}^{t+1}$ to the set $\{\mathcal{T} \in \mathbb{R}^{n_1\times n_2\times n_3} | \|\mathcal{T}\|_1 \le \mu\}$ using the algorithm detailed in [54].
7: **end for**
8: **return** $\hat{\mathcal{L}} = \mathcal{L}^{t+1}$ and $\hat{\mathcal{C}} = \mathcal{C}^{t+1}$.

C. Comparison of Two Algorithms

Though TF-QRPCA and TNN-QPRCA are both applications of PALM, they have significant differences in the following aspects. 1) TF-QRPCA uses tensor factorization to approximate the low-tubal-rank tensor, then avoids conducting the t-SVD in each iteration as we do in TNN-QRPCA. For this reason, TF-QRPCA has relatively lower computational cost than TNN-QRPCA. 2) Compared with TNN-QRPCA, TF-QRPCA relies on a more strict low-rankness condition. Though we have provided a method to estimate the tubal rank, the low-rankness is usually corrupted by sparse and random noises, making it hard to obtain a satisfying estimate in practice. In addition, many practical data do not strictly

satisfy the low-rankness hypothesis. Figure 1 gives an example to show that a color image can be regarded as an approximately low-tubal-rank tensor that has a relatively small tensor nuclear norm. But it does not strictly satisfy the low-rankness hypothesis since many singular values are not zero though they are very small. On the contrary, TNN-QRPCA is based on a weaker low-rankness condition, i.e., relatively small tensor nuclear norm, which is more easily satisfied. Therefore, TNN-QRPCA is usually more robust to sparse noises than TF-QRPCA, which is supported by the experimental results in Section VI.



(a) (b)

Fig. 1: Color images can be regarded as approximately low-tubal-rank tensors. (a) A color image can be modeled as a tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{512\times512\times3}$; (b) plot of the natural logarithm of singular values of $\boldsymbol{\mathcal{M}}$.

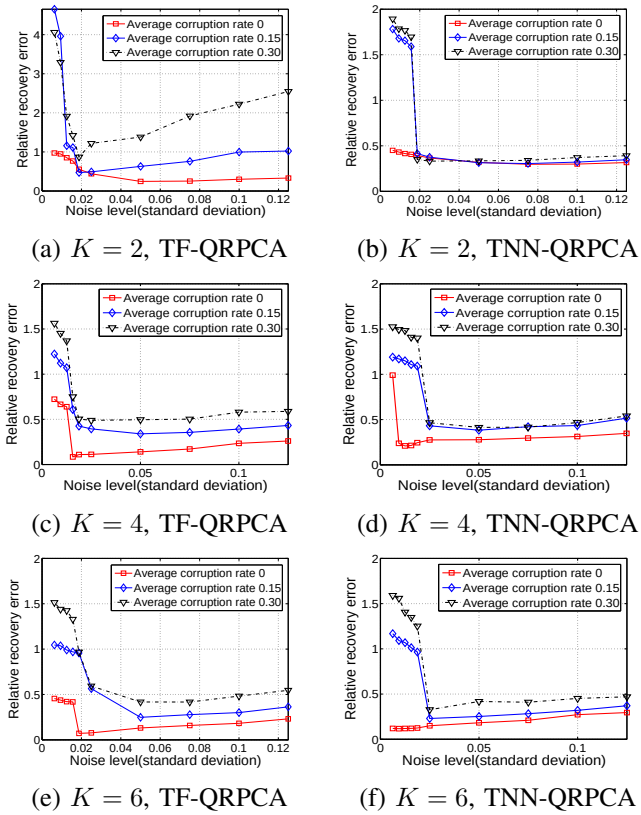
VI. Simulations

In this section, we conduct a series of numerical experiments on both synthetic and real-world data to test the validity of the proposed algorithms and the correctness of the theoretical analysis. We perform all experiments on a 64-bit PC with a Gold-5122 3.60GHz CPU and 192 GB memory. The relative recovery error is defined as $\|\widetilde{\boldsymbol{\mathcal{L}}} - \hat{\boldsymbol{\mathcal{L}}}\|_F / \|\widetilde{\boldsymbol{\mathcal{L}}}\|_F$, where $\widetilde{\boldsymbol{\mathcal{L}}}$ denotes the actual tensor and $\hat{\boldsymbol{\mathcal{L}}}$ is the recovered tensor. The corruption rate $s/(n_1n_2n_3)$ denotes the proportion of nonzero entries in $\widetilde{\boldsymbol{\mathcal{C}}}$. In our synthetic experiments, we generate the actual tensor $\widetilde{\boldsymbol{\mathcal{L}}}$ with tubal rank r as a product $\widetilde{\boldsymbol{\mathcal{L}}} = \boldsymbol{\mathcal{P}} * \boldsymbol{\mathcal{Q}}$, where $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{n_1\times r\times n_3}$ and $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{r\times n_2\times n_3}$ are both with entries drawn i.i.d. from the uniform distribution on $[-0.5, 0.5]$. Then we scale it so that $\|\widetilde{\boldsymbol{\mathcal{L}}}\|_\infty = 1$. Entries of the sparse component $\widetilde{\boldsymbol{\mathcal{C}}}$ are drawn i.i.d. from the uniform distribution on $[-1, 1]$. The entries of $\boldsymbol{\mathcal{N}}$ are drawn i.i.d. from the logistic distribution $g(x) = 1/(1 + e^{-x/\sigma})$.

A. Synthetic Experiments

We first conduct an experiment to determine an appropriate choice for σ so that our algorithms perform well with the dither tensor $\boldsymbol{\mathcal{N}}$. We fix the tensor size $(n_1, n_2, n_3) = (100, 100, 10)$ and the tubal rank $r = 2$. We consider three quantization levels $K = 2, 4, 6$ and choose quantization boundaries $\{-\infty, 0, +\infty\}$ for $K = 2$, $\{-\infty, -0.5, 0, 0.5, +\infty\}$ for $K = 4$, and $\{-\infty, -2/3, -1/3, 0, 1/3, 2/3, +\infty\}$ for $K = 6$. The measurements are observed via (15) with $W = 1$. The experimental results are plotted in Fig. 2. We can easily find that, for both approaches, the recovery performance deteriorates when there is too little noise (when σ is small) and when there is too much noise (when σ is large). In the regime where the noise is of moderate power, we observe better performance for both approaches. Consider $K = 6$; $\sigma = 0.025$ and $\sigma = 0.05$ are respectively proper choices for TF-QRPCA and TNN-QRPCA, where both approaches achieve relatively small errors when the observations are corrupted by sparse noises.

From Fig. 2, we observe that TNN-QRPCA seems to be more robust to sparse noises than TF-QRPCA. To further explore how sparse noise affects the recovery performance of the proposed algorithms, we conduct an additional experiment. Following the same setting as the first experiment, we change the corruption rates to test the relative recovery error and report the average results in Fig. 3 over 100 runs. From these results, we find that TF-QRPCA performs better than TNN-QRPCA in the low-corruption rate region. However, TNN-QRPCA is more robust to sparse noises in the high-corruption rate region especially when the quantization level is low, e.g., $K = 2, 4$, as compared with TF-QRPCA. In fact, as we noted in Section V-C, the advantage of TF-QRPCA in the low-corruption rate region comes from a more strict low-rankness penalty. When corrupted by uniform sparse noises, the low-rankness is severely damaged, which significantly reduces the recovery performance of TF-QRPCA.

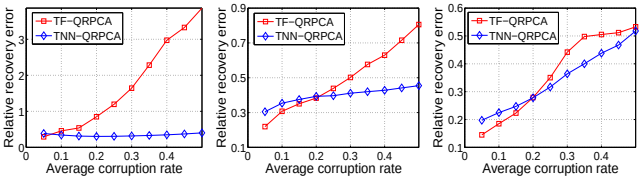


(a) $K = 2$, TF-QRPCA (b) $K = 2$, TNN-QRPCA

(c) $K = 4$, TF-QRPCA (d) $K = 4$, TNN-QRPCA

(e) $K = 6$, TF-QRPCA (f) $K = 6$, TNN-QRPCA

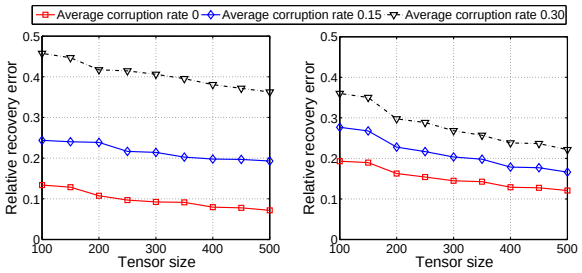
Fig. 2: Average relative recovery errors in the results obtained using TF-QRPCA and TNN-QRPCA over a range of different noise levels (different σ in the logistic distribution $g(x) = 1/(1 + e^{-x/\sigma})$) under different quantization levels.



(a) $K = 2$ (b) $K = 4$ (c) $K = 6$

Fig. 3: Comparison of the recovery performance of TF-QRPCA and TNN-QRPCA under different corruption rates and different quantization levels. We observe a more robust performance of TNN-QRPCA than TF-QRPCA in the high-corruption rate region.

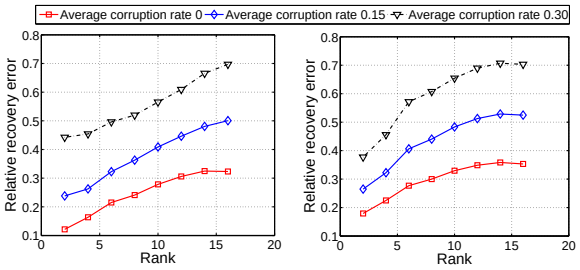
In the third experiment, we test the performance of the proposed algorithms when the tensor size and tubal rank change, respectively. We choose quantization level $K = 6$ with quantization boundaries $\{-\infty, -2/3, -1/3, 0, 1/3, 2/3, +\infty\}$ and consider three corruption rates, i.e., 0, 0.20, and 0.40. First, we change n_1, n_2 when fixing $n_3 = 20$, $r = 4$ and report the results in Fig. 4 over 100 runs. We observe that the recovery errors obtained by both algorithms decrease with increased n_1 and n_2, which is in line with our theoretical analysis. Second, we change the tubal rank from 2 to 16 when fixing $n_1 \times n_2 \times n_3 = 100 \times 100 \times 20$ and show the averaged results in Fig. 5 over 100 trials. We observe that the relative recovery errors increase when the tubal rank is increased, which is also consistent with our theoretical results.



(a) TF-QRPCA (b) TNN-QRPCA

Fig. 4: Average relative recovery errors ($K = 6$) in the results obtained using TF-QRPCA and TNN-QRPCA on tensors of different scales. With the increase of the tensor scale, we observe a decaying trend in the relative recovery error.

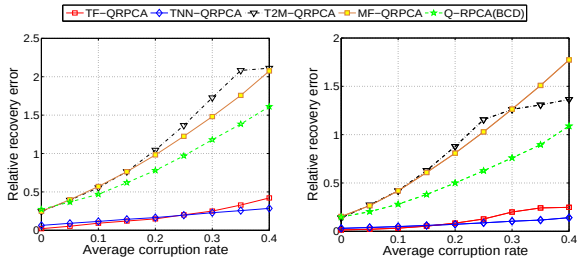
Next, we test the effect of quantization level and copies W of repeated measurements on the recovery performance of both algorithms. We keep the tensor size $n_1 = n_2 = 100$, $n_3 = 20$, and the tubal rank $r = 2$ and consider quantization levels $K = 2, 4, 6, 8$ and copies of repeated observations $W = 1, 2, 4$. Quantization boundaries for $K = 2, 4, 6$ are the same as the above experiments and for $K = 8$ are $\{-\infty, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, +\infty\}$. The averaged relative recovery errors are reported in TABLE I over 100 runs. In general, the recovery performances of both algorithms improve with the quantization level K increasing. Additionally, we find that TNN-QRPCA performs extremely robust



(a) TF-QRPCA (b) TNN-QRPCA

Fig. 5: Average relative recovery errors ($K = 6$) in the results obtained using TF-QRPCA and TNN-QRPCA over a range of different tubal ranks. With the increase of the tensor tubal rank, we observe an increasing trend in the relative recovery error.

when $K = 2$, which obtains a better recovery performance than the results obtained when $K = 4$. This may be due to the robustness of 1-bit quantization which only preserves the signs of observations [56], [57]. From the experimental results, we also conclude that the recovery performances of both algorithms are improved with the increase of W when fixing K, which is also consistent with our theoretical analysis. Finally, we compare the proposed algorithms, TF-QRPCA and TNN-QRPCA, with MF-QRPCA [7], Q-RPCA(BCD) [9], and T2M-QRPCA [39]. We evaluate the recovery performance by calculating the relative recovery errors. The corrupted rate by sparse noises varies from 0 to 40% and the results (averaged over 100 runs) are reported in Figure 6. It can be easily seen that the proposed algorithms give a significantly improved performance under all corruption rates for low-tubal-rank tensor data, as compared with the other methods.



(a) $K = 4$ (b) $K = 6$

Fig. 6: Comparison of the recovery performance of different algorithms under different corruption rates of sparse noises. We see a significantly improved performance of the proposed algorithms.

B. Multispectral Image Restoration

In this part, we apply our methods to multispectral image restoration from quantized and sparsely corrupted observations since multispectral data possess the typical tensor form. We use Stuffed TOY data to evaluate the proposed methods. The Stuffed TOY data come from the CAVE Multispectral

TABLE I: Comparison of the recovery performance of TF-QRPCA and TNN-QRPCA under different quantization levels and different copies of repeated observations.

$p=0$												
Methods	K=2			K=4			K=6			K=8		
	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$
TF-QRPCA	0.2525	0.1586	0.0973	0.1400	0.1018	0.0723	0.1173	0.0829	0.0589	0.1041	0.0746	0.0530
TNN-QRPCA	0.2956	0.2724	0.2418	0.2502	0.1830	0.1382	0.1742	0.1295	0.1142	0.1572	0.1209	0.1110
$p=0.2$												
Methods	K=2			K=4			K=6			K=8		
	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$
TF-QRPCA	0.8202	0.4265	0.4204	0.3839	0.3223	0.2647	0.2878	0.2070	0.1298	0.2132	0.1577	0.1092
TNN-QRPCA	0.3070	0.2789	0.2602	0.3852	0.2815	0.2401	0.2814	0.2159	0.1626	0.2440	0.1880	0.1440
$p=0.4$												
Methods	K=2			K=4			K=6			K=8		
	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$	$W=1$	$W=2$	$W=4$
TF-QRPCA	1.8099	0.6231	0.5027	0.4957	0.3877	0.3462	0.4322	0.2791	0.2117	0.2883	0.2159	0.1524
TNN-QRPCA	0.3178	0.2801	0.2647	0.4098	0.2943	0.2820	0.3226	0.2646	0.1945	0.2905	0.2429	0.1905

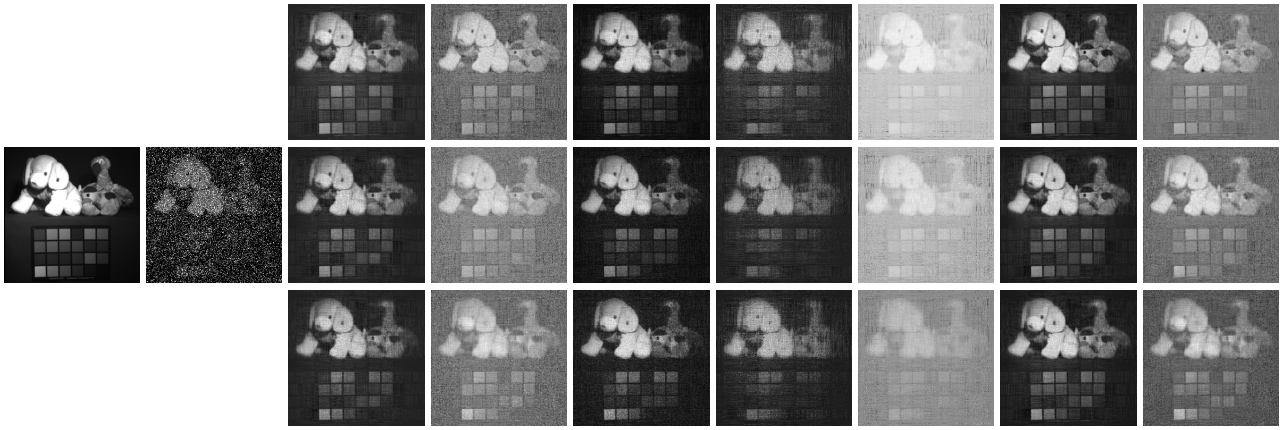
Image Database [58], which consists of 31 spectral frames with each of size 512×512. We resize the image to 31 spectral frames with each of size 256×256 and then normalize all pixels in the range $[0, 1]$ via the max-min formula to conduct our experiments. The random noises are drawn i.i.d. from the logistic distribution $g(x) = 1/(1 + e^{-x/\sigma})$ with $\sigma = 0.05$. We fix the quantization level $K = 6$, empirically set the tubal rank $r = 24$, and choose the quantization boundaries $\{-\infty, 1/6, 1/3, 1/2, 2/3, 5/6, +\infty\}$. We compare the proposed methods TNN-QRPCA and TF-QRPCA with MF-QRPCA, Q-RPCA(BCD), and T2M-QRPCA under three corruption rates $p = 0, 15\%, 30\%$. To illustrate that repeated measurements can significantly improve the recovery performance of TF-QRPCA and TNN-QRPCA, we also apply them to repeated measurements (15) with copies $W = 4$ (denoted as TF-QRPCA(W=4) and TNN-QRPCA(W=4), respectively). For MF-QRPCA and Q-RPCA(BCD), we apply them on each frame separately and combine the obtained images to obtain the recovered result. We evaluate the recovery performance using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) indices [59]. For an $n_1 \times n_2 \times n_3$ image $\boldsymbol{\mathcal{C}}$ and its recovered image $\hat{\boldsymbol{\mathcal{C}}}$, PSNR is defined as $\mathrm{PSNR} = 10 \log_{10} \left(\frac{n_1 \times n_2 \times n_3 \|\boldsymbol{\mathcal{C}}\|_\infty^2}{\|\hat{\boldsymbol{\mathcal{C}}} - \boldsymbol{\mathcal{C}}\|_F^2} \right)$; SSIM is defined as

$$\mathrm{SSIM} = \frac{\left(2\mu_{\boldsymbol{\mathcal{C}}}\mu_{\hat{\boldsymbol{\mathcal{C}}}} + (0.01\bar{\omega})^2\right)\left(2\sigma_{\boldsymbol{\mathcal{C}},\hat{\boldsymbol{\mathcal{C}}}} + (0.03\bar{\omega})^2\right)}{\left(\mu_{\boldsymbol{\mathcal{C}}}^2 + \mu_{\hat{\boldsymbol{\mathcal{C}}}}^2 + (0.01\bar{\omega})^2\right)\left(\sigma_{\boldsymbol{\mathcal{C}}}^2 + \sigma_{\hat{\boldsymbol{\mathcal{C}}}}^2 + (0.03\bar{\omega})^2\right)},$$

where $\mu_{\boldsymbol{\mathcal{C}}}$, $\mu_{\hat{\boldsymbol{\mathcal{C}}}}$, $\sigma_{\boldsymbol{\mathcal{C}}}$, $\sigma_{\hat{\boldsymbol{\mathcal{C}}}}$, $\sigma_{\boldsymbol{\mathcal{C}},\hat{\boldsymbol{\mathcal{C}}}}$, and $\bar{\omega}$ are the local means, standard deviation, cross-covariance, and dynamic range of the pixel values for images $\boldsymbol{\mathcal{C}}$, $\hat{\boldsymbol{\mathcal{C}}}$. PSNR values and SSIM values are shown in TABLE II. We also plot the first spectral frame and correspondingly recovered results by different methods in Fig 7. From these results, we have the following observations. First, the proposed TNN-QRPCA gives a significantly improved performance especially when the observations are corrupted, as compared with other competing methods. Second, TF-QRPCA and MF-QRPCA attain relatively worse performance. The reason may be that the two methods are based on tensor (matrix) factorization, which depends on a strict low-rankness hypothesis. On the contrary, the Stuffed TOY data, which has many small but nonzero singular values, do not strictly satisfy the low-rankness condition. Even so, TF-QRPCA still shows obvious advantages over MF-QRPCA, which demonstrates the power of tensor modeling techniques. Third, we observe that adding the copies of repeated measurements can effectively improve the recovery performance of TNN-QRPCA and TF-QRPCA, which validates our theoretical results.

C. Color Image Denoising

We now test the recovery performance of the proposed algorithms on color image data. We use the Berkeley Segmentation Dataset [60] for the test. We randomly draw 25 images from the Berkeley Segmentation Dataset where each image consists of 3 channels with each of size 481×321. Pixels of each image are normalized in the range $[0, 1]$. We follow the same setting about the random noises and quantization boundaries as the previous experiment. We compare the proposed algorithms, TNN-QRPCA, TF-QRPCA, TNN-QRPCA(W=4), and TF-QRPCA(W=4), with MF-QRPCA, Q-RPCA(BCD), and T2M-QRPCA under corruption rates $p = 0$ and $p = 20\%$, respectively. Fig. 8 plots the comparisons of the PSNR values and SSIM values obtained by these methods on all 25 images when $p = 0$. We also show some examples and correspondingly recovered results in Fig. 9. Fig. 10 plots the comparison of the PSNR values and SSIM values obtained by different algorithms on all 25 images when $p = 20\%$ with some visual examples and correspondingly recovered results shown in Fig. 11. Especially, in Fig. 11 (k), we show the ratio of PSNR values and SSIM values obtained when $p = 20\%$ to that obtained when $p = 0$ based on these examples. It is easy to find that, in both corruption rates, TNN-QRPCA has a better performance than other methods no matter in quantitative indices or visual results in most cases. Besides, from Fig. 11 (k), when the observations are corrupted by sparse noises, our methods, TNN-QRPCA and TF-QRPCA, lose less information in both PSNR values and SSIM values than MF-QRPCA, Q-RPCA(BCD), and T2M-QRPCA, which reflect better robustness of the proposed methods.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 7: The first spectral frame of Stuffed TOY data and correspondingly recovered results. (a) Original image; (b) observed image; (c) recovered images by TNN-QRPCA; (d) recovered images by TF-QRPCA; (e) recovered images by T2M-QRPCA; (f) recovered images by Q-RPCA(BCD); (g) recovered images by MF-QRPCA; (h) recovered images by TNN-QRPCA(W=4); (i) recovered images by TF-QRPCA(W=4).

TABLE II: Comparison of the PSNR values and SSIM values obtained by different methods on Stuffed TOY data. Here $n_1 = 256$, $n_2 = 256$, $n_3 = 31$. Three corruption rates, $p = 0$, $p = 15\%$, and $p = 30\%$, are considered, respectively.

Methods	$p = 0$		$p = 15\%$		$p = 30\%$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TNN-QRPCA	29.8676	0.7472	28.2540	0.7061	26.3275	0.6644
TF-QRPCA	22.1195	0.2697	20.4014	0.2711	20.1522	0.2699
T2M-QRPCA	29.5728	0.7569	25.3860	0.5006	23.0276	0.4123
Q-RPCA(BCD)	26.7770	0.5527	25.7516	0.5825	24.5451	0.5582
MF-QRPCA	18.7209	0.2911	17.7167	0.2542	18.4769	0.2286
TNN-QRPCA(W=4)	**33.3797**	**0.8349**	**30.6875**	**0.7845**	**28.3931**	**0.7566**
TF-QRPCA(W=4)	27.6940	0.5454	24.0461	0.4277	22.6417	0.3627

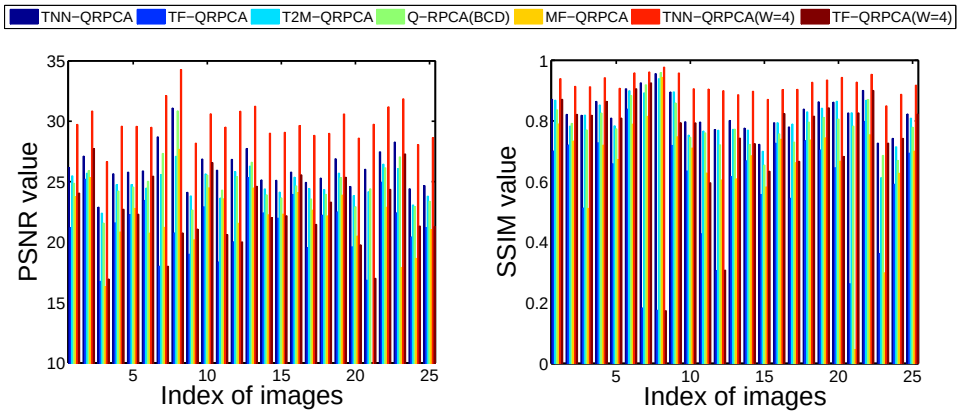


Fig. 8: Comparison of the PSNR values and SSIM values obtained by TNN-QRPCA, TF-QRPCA, T2M-QRPCA, Q-RPCA(BCD), MF-QRPCA, TNN-QRPCA(W=4), and TF-QRPCA(W=4) on 25 tested images when $p = 0$.

VII. Conclusion and Future Work

This work provides the first theoretical study of quantized tensor robust principal component analysis which aims to recover a low-rank tensor from noisy, quantized, and sparsely corrupted measurements. We propose two optimization problems and prove that under mild assumptions, we can recover both the low-rank and the sparse tensors approximately by solving the proposed problems. We also propose two proximal algorithms to solve the proposed problems, respectively. Benefiting from the properties of t-SVD, our methods and theoretical results are both natural generalizations of quantized matrix robust principal component analysis. Finally, numerical experiments verify our theoretical results and experiments with multispectral image recovery and color image denoising

(a) (b) (c) (d) (e) (f) (g) (h)

Index	1		2		3		4		5	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TNN-QRPCA	26.17	0.87	27.11	0.82	25.89	0.91	27.48	0.90	26.90	0.86
TF-QRPCA	21.21	0.70	25.34	0.73	23.45	0.84	24.95	0.80	23.89	0.74
T2M-QRPCA	25.49	0.87	25.70	0.78	24.47	0.90	26.45	0.87	25.71	0.84
Q-RPCA(BCD)	24.87	0.84	25.92	0.79	20.03	0.88	26.16	0.87	25.34	0.81
MF-QRPCA	23.77	0.79	25.20	0.72	20.74	0.79	22.88	0.75	22.50	0.71
TNN-QRPCA(W=4)	**29.74**	**0.94**	**30.85**	**0.91**	**29.50**	**0.96**	**31.20**	**0.95**	**30.61**	**0.94**
TF-QRPCA(W=4)	24.06	0.87	27.73	0.82	25.46	0.91	24.37	0.90	25.37	0.84

(i) Comparison of the PSNR values and the SSIM values on the above 5 images.

Fig. 9: Recovery performance comparison on 5 example images when the quantized measurements are without sparse noises. (a) Original images; (b) recovered images by TNN-QRPCA; (c) recovered images by TF-QRPCA; (d) recovered images by T2M-QRPCA; (e) recovered images by Q-RPCA(BCD); (f) recovered images by MF-QRPCA; (g) recovered images by TNN-QRPCA(W=4); (h) recovered images by TF-QRPCA(W=4); (i) show the comparison of the PSNR values and SSIM values on the above 5 images.

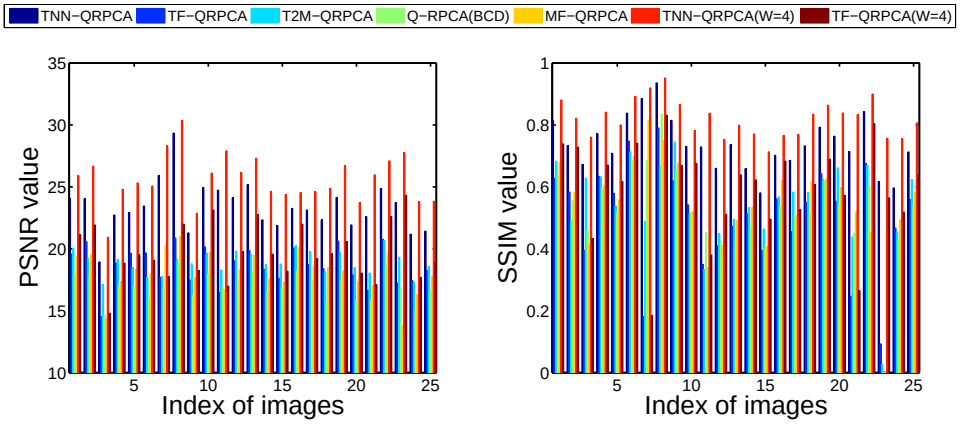


Fig. 10: Comparison of the PSNR values and SSIM values obtained by TNN-QRPCA, TF-QRPCA, T2M-QRPCA, Q-RPCA(BCD), MF-QRPCA, TNN-QRPCA(W=4), and TF-QRPCA(W=4) on 25 tested images when $p = 20\%$.

demonstrate the effectiveness of the proposed algorithms. It is worth mentioning that this work focuses on the analysis for 3-way tensors. But the established results are extendable to the case of order-$p(p > 3)$ tensors, by employing the t-SVD for order-p tensors in [61] and the order-p t-SVD rank and the order-p TNN in [62], since the loss function, $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$,

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Index	1		2		3		4		5	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TNN-QRPCA	24.07	0.81	24.05	0.73	23.46	0.84	24.87	0.84	24.14	0.79
TF-QRPCA	19.57	0.63	20.57	0.58	19.68	0.75	20.76	0.67	20.63	0.64
T2M-QRPCA	20.05	0.68	19.23	0.49	17.66	0.71	20.66	0.67	19.73	0.62
Q-RPCA(BCD)	17.94	0.64	17.90	0.55	16.10	0.68	18.65	0.60	17.51	0.62
MF-QRPCA	19.42	0.68	19.55	0.58	17.99	0.70	19.67	0.63	18.22	0.45
TNN-QRPCA(W=4)	**25.92**	**0.88**	**26.69**	**0.82**	**25.08**	**0.89**	**27.11**	**0.90**	**26.71**	**0.86**
TF-QRPCA(W=4)	21.13	0.74	21.94	0.73	19.09	0.74	22.62	0.81	20.58	0.69

(j) Comparison of the PSNR values and the SSIM values on the above 5 images.

Index	1		2		3		4		5	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TNN-QRPCA	92.0%	93.1%	**88.7%**	89.0%	**90.6%**	92.3%	90.5%	93.3%	**89.7%**	**91.9%**
TF-QRPCA	**92.3%**	90.0%	81.8%	79.5%	83.9%	89.3%	83.2%	83.8%	86.4%	86.5%
T2M-QRPCA	78.7%	78.2%	74.8%	62.8%	72.2%	78.9%	78.1%	77.0%	74.8%	73.8%
Q-RPCA(BCD)	72.1%	76.2%	69.1%	69.6%	80.4%	77.3%	71.3%	70.0%	69.1%	76.5%
MF-QRPCA	81.7%	86.1%	77.6%	80.6%	86.7%	88.6%	86.0%	84.0%	81.0%	63.4%
TNN-QRPCA(W=4)	87.2%	**93.6%**	86.5%	**90.1%**	85.0%	**92.7%**	86.9%	**94.7%**	87.3%	91.5%
TF-QRPCA(W=4)	87.8%	85.1%	79.1%	89.0%	75.0%	81.3%	**92.8%**	90.0%	81.1%	82.1%

(k) The ratio of PSNR values and SSIM values obtained when $p = 20\%$ to that obtained when $p = 0$.

Fig. 11: Recovery performance comparison on 5 example images when the quantized measurements are with 20% corruption. (a) Original images; (b) observed images; (c) recovered images by TNN-QRPCA; (d) recovered images by TF-QRPCA; (e) recovered images by T2M-QRPCA; (f) recovered images by Q-RPCA(BCD); (g) recovered images by MF-QRPCA; (h) recovered images by TNN-QRPCA(W=4); (i) recovered images by TF-QRPCA(W=4); (j) show the comparison of the PSNR values and SSIM values on the above 5 images; (k) list the ratio of the PSNR values and SSIM values obtained when $p = 20\%$ to that obtained when $p = 0$.

applies componentwise on $\boldsymbol{\mathcal{X}}$ and the definitions of TNN and t-SVD rank (tubal rank) for 3-order tensors coincide with the higher-order case.

This work supposes that the quantization boundaries are known to ease our analysis. We also note that the recent work [45] has provided an adaptive method for the unknown quantization boundaries. In their method, the unknown k-th order tensor $\boldsymbol{\mathcal{X}}$ and unknown quantization boundaries $\boldsymbol{\omega} = \{\omega_0, \cdots, \omega_K\}$ are stacked to a new $k+1$-th order tensor, all regarded as the parameters of the log-likelihood term $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}, \boldsymbol{\omega})$. Then, they are alternatively optimized while holding others fixed. Following the idea, the proposed algorithms in this paper have the potential to be generalized for the unknown quantization since they are also based on alternating procedures. However, the analysis for the statistical error depends on some different assumptions and technologies and we leave it for future work.

APPENDIX

In this section, we provide proofs for Theorem 1, Theorem 2, Theorem 3, and Theorem 4, respectively.

A. Proof of Theorem 1

Before beginning the proof, we define $\hat{\boldsymbol{\theta}} = \mathrm{vec}(\hat{\boldsymbol{\mathcal{X}}}) \in \mathbb{R}^{n_1n_2n_3}$, $\widetilde{\boldsymbol{\theta}} = \mathrm{vec}(\widetilde{\boldsymbol{\mathcal{X}}}) \in \mathbb{R}^{n_1n_2n_3}$, $\mathscr{F}_{\mathscr{Y}}(\hat{\boldsymbol{\theta}}) = F_{\mathscr{Y}}(\hat{\boldsymbol{\mathcal{X}}})$, and $\mathscr{F}_{\mathscr{Y}}(\widetilde{\boldsymbol{\theta}}) = F_{\mathscr{Y}}(\widetilde{\boldsymbol{\mathcal{X}}})$. Applying a second-order Taylor expansion implies

$$\begin{aligned}\mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta}) =& \mathscr{F}_{\mathscr{Y}}(\widetilde{\boldsymbol{\theta}}) + \langle \nabla_{\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\widetilde{\boldsymbol{\theta}}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}} \rangle \\ &+ \frac{1}{2} \langle \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}, \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta}^{\#})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \rangle,\end{aligned} \tag{57}$$

where $\boldsymbol{\theta}^{\#} = \widetilde{\boldsymbol{\theta}} + \eta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$ for some $\eta \in [0,1]$, and thus corresponding tensor $\boldsymbol{\mathcal{X}}^{\#} = \widetilde{\boldsymbol{\mathcal{X}}} + \eta(\boldsymbol{\mathcal{X}} - \widetilde{\boldsymbol{\mathcal{X}}})$.

Lemma 1. *Suppose $(\boldsymbol{\mathcal{L}}', \boldsymbol{\mathcal{C}}'), (\boldsymbol{\mathcal{L}}'', \boldsymbol{\mathcal{C}}'') \in \mathbb{K}_{r,s}$. Let $\boldsymbol{\mathcal{X}}' = \boldsymbol{\mathcal{L}}' + \boldsymbol{\mathcal{C}}'$, $\boldsymbol{\mathcal{X}}'' = \boldsymbol{\mathcal{L}}'' + \boldsymbol{\mathcal{C}}''$, $\boldsymbol{\theta}' = \mathrm{vec}(\boldsymbol{\mathcal{X}}')$, $\boldsymbol{\theta}'' = \mathrm{vec}(\boldsymbol{\mathcal{X}}'')$, and $d = (n_1 + n_2)n_3$. Then with probability at least $1 - 1/d$, we have*

$$\begin{aligned}&|\langle \nabla_{\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta}'), \boldsymbol{\theta}'' - \boldsymbol{\theta}' \rangle| \\ \le& 2c^* L_\alpha \sqrt{rWn_{(2)}n_3 \log(d)} (\|\boldsymbol{\mathcal{X}}'' - \boldsymbol{\mathcal{X}}'\|_F + 2\alpha\sqrt{2s}) \\ &+ 4W\alpha s L_\alpha,\end{aligned}$$

where $c^* := 1 + \sqrt{3}$.

Proof. The proof relies on the following proposition.

Proposition 4 (See [15])**.** *Consider a finite sequence of independent random matrices $\{\boldsymbol{Z}_i\}_{1 \le i \le n} \in \mathbb{R}^{n_1 \times n_2}$ that satisfy $\mathbb{E}(\boldsymbol{Z}_i) = \boldsymbol{0}$ and for some $U > 0$, $\|\boldsymbol{Z}_i\| \le U$ for all $i \in [n]$. Then for $d = n_1 + n_2$ and any $t > 0$, we have*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \right\| > t \right) \le d \exp\left(-\frac{nt^2/2}{\sigma_Z^2 + Ut/3} \right),$$

where

$$\sigma_Z^2 = \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\boldsymbol{Z}_i \boldsymbol{Z}_i^T] \right\|, \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\boldsymbol{Z}_i^T \boldsymbol{Z}_i] \right\| \right\}.$$

In particular, it implies that with probability at least $1 - \exp(-t)$, it holds that

$$\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \right\| \le c^* \max\left\{ \sigma_Z \sqrt{\frac{t + \log(d)}{n}}, \frac{U(t + \log(d))}{3n} \right\},$$

where $c^* := 1 + \sqrt{3}$.

Now, consider

$$\begin{aligned}\mathcal{Z}_{ijk} &= [-L_\alpha^{-1} W^{-1} \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}')]_{ijk} \\ &= -L_\alpha^{-1} W^{-1} \sum_{t=1}^W \sum_{l=1}^K \frac{\dot{f}_l(\mathcal{X}'_{ijk})}{f_l(\mathcal{X}'_{ijk})} \mathbb{1}_{\{\mathcal{Y}^t_{ijk} = l\}},\end{aligned}$$

and let $\boldsymbol{\mathcal{Z}}_{ijk}$ denote a tensor of size $n_1 \times n_2 \times n_3$ whose (i,j,k)-th entry is $\mathcal{Z}_{ijk}$ and other entries are all zero. Then we need to bound $\|\nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}')\| = L_\alpha W \| \sum_{i,j,k} \boldsymbol{\mathcal{Z}}_{ijk} \|$. Recalling

$$\left\| \sum_{i,j,k} \boldsymbol{\mathcal{Z}}_{ijk} \right\| = \left\| \sum_{i,j,k} \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}) \right\|,$$

we next turn to leverage Proposition 4 to bound the term $\| \sum_{i,j,k} \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}) \|$. It is obvious that $\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})$ are independent from each other. Using (9) and the fact that $\sum_{l=1}^K f_l(x) = 1$, one can check that $\mathbb{E}(\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})) = \boldsymbol{0}$ and $\|\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})\| \le 1$. Let $\boldsymbol{E}_{qq}$ denote a matrix of size $n_1n_3 \times n_1n_3$ whose (q,q)-entry is 1 and other entries are 0. We have

$$\begin{aligned}&\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}) \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T = -L_\alpha^{-2} W^{-2} \\ &\cdot \sum_{t=1}^W \sum_{l=1}^K \mathbb{1}_{\{\mathcal{Y}^t_{ijk} = l\}} \frac{\dot{f}_l^2(\mathcal{X}'_{ijk})}{f_l^2(\mathcal{X}'_{ijk})} \sum_{q=0}^{n_3 - 1} \boldsymbol{E}_{(qn_1+i)(qn_1+i)},\end{aligned}$$

and thus

$$\begin{aligned}&\left\| \sum_{i=1}^{n_1} \mathbb{E}[\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}) \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T] \right\| \\ \le& \frac{1}{W} \left\| \sum_{i=1}^{n_1} \sum_{q=0}^{n_3 - 1} \boldsymbol{E}_{(qn_1+i)(qn_1+i)} \right\| = \frac{1}{W}.\end{aligned}$$

Therefore, we have

$$\left\| \frac{1}{n_1 n_2 n_3} \sum_{i,j,k} \mathbb{E}[\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}) \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T] \right\| \le \frac{1}{Wn_1}.$$

Similarly,

$$\left\| \frac{1}{n_1 n_2 n_3} \sum_{i,j,k} \mathbb{E}[\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T \mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})] \right\| \le \frac{1}{Wn_2}.$$

Applying Proposition 4 with $n = n_1n_2n_3$, $U = 1$, $\sigma_Z^2 = 1/(Wn_{(2)})$, $d = (n_1 + n_2)n_3$, and $t = \log(d)$ implies that with probability at least $1 - 1/d$,

$$\left\| \sum_{i,j,k} \boldsymbol{\mathcal{Z}}_{ijk} \right\| \le c^* \max\left\{ \sqrt{\frac{2n_{(1)}n_3 \log(d)}{W}}, \frac{2\log(d)}{3} \right\}.$$

Note that

$$\begin{aligned}\|\boldsymbol{\mathcal{L}}'' - \boldsymbol{\mathcal{L}}'\|_* &= \frac{1}{n_3} \|\overline{\boldsymbol{L}''} - \overline{\boldsymbol{L}'}\|_* \le \frac{1}{n_3} \sqrt{2rn_3} \|\overline{\boldsymbol{L}''} - \overline{\boldsymbol{L}'}\|_F \\ &= \sqrt{2r} \|\boldsymbol{\mathcal{L}}'' - \boldsymbol{\mathcal{L}}'\|_F,\end{aligned}$$

and $\sqrt{2n_{(1)}n_3 \log(d)/W}$ is much larger than $2\log(d)$. We then obtain that with probability at least $1 - 1/d$

$$\|\nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}')\| \|\boldsymbol{\mathcal{L}}'' - \boldsymbol{\mathcal{L}}'\|_* \le 2c^* L_\alpha W h \|\boldsymbol{\mathcal{L}}'' - \boldsymbol{\mathcal{L}}'\|_F,$$

where $h := \sqrt{rn_{(1)}n_3 \log(d)/W}$. Meanwhile, we also have

$$\langle \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}'), \boldsymbol{\mathcal{C}}'' - \boldsymbol{\mathcal{C}}' \rangle \le 4W\alpha s L_\alpha$$

by the assumption (9), the fact that $\|\boldsymbol{\mathcal{C}}''-\boldsymbol{\mathcal{C}}'\|_\infty \le 2\alpha$, and $\|\boldsymbol{\mathcal{C}}''-\boldsymbol{\mathcal{C}}'\|_0 \le 2s$. Therefore, we have that

$$\begin{aligned}
&|\langle \nabla_{\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta}'), \boldsymbol{\theta}''-\boldsymbol{\theta}'\rangle| \\
&= |\langle \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}'), \boldsymbol{\mathcal{X}}''-\boldsymbol{\mathcal{X}}'\rangle| \\
&\le |\langle \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}'), \boldsymbol{\mathcal{L}}''-\boldsymbol{\mathcal{L}}'\rangle| + |\langle \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}'), \boldsymbol{\mathcal{C}}''-\boldsymbol{\mathcal{C}}'\rangle| \\
&\le \|\nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}')\| \|\boldsymbol{\mathcal{L}}''-\boldsymbol{\mathcal{L}}'\|_* + |\langle \nabla_{\boldsymbol{\mathcal{X}}} F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}'), \boldsymbol{\mathcal{C}}''-\boldsymbol{\mathcal{C}}'\rangle| \\
&\le 2c^* W L_\alpha h \|\boldsymbol{\mathcal{L}}''-\boldsymbol{\mathcal{L}}'\|_F + 4W\alpha s L_\alpha \\
&\le 2c^* W L_\alpha h (\|\boldsymbol{\mathcal{X}}''-\boldsymbol{\mathcal{X}}'\|_F + \|\boldsymbol{\mathcal{C}}''-\boldsymbol{\mathcal{C}}'\|_F) + 4W\alpha s L_\alpha \\
&\le 2c^* W L_\alpha h (\|\boldsymbol{\mathcal{X}}''-\boldsymbol{\mathcal{X}}'\|_F + 2\alpha\sqrt{2s}) + 4W\alpha s L_\alpha
\end{aligned}$$

holds with probability at least $1-1/d$. □

Lemma 2. *Suppose $(\boldsymbol{\mathcal{L}}', \boldsymbol{\mathcal{C}}'), (\boldsymbol{\mathcal{L}}'', \boldsymbol{\mathcal{C}}'') \in \mathbb{K}_{r,s}$. Let $\boldsymbol{\mathcal{X}}' = \boldsymbol{\mathcal{L}}' + \boldsymbol{\mathcal{C}}'$, $\boldsymbol{\mathcal{X}}'' = \boldsymbol{\mathcal{L}}'' + \boldsymbol{\mathcal{C}}''$, $\boldsymbol{\theta}' = \mathrm{vec}(\boldsymbol{\mathcal{X}}')$, and $\boldsymbol{\theta}'' = \mathrm{vec}(\boldsymbol{\mathcal{X}}'')$. Then for any $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}' + \eta(\boldsymbol{\theta}''-\boldsymbol{\theta}')$ and any $\eta \in [0,1]$, we have*

$$\langle \boldsymbol{\theta}''-\boldsymbol{\theta}', (\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta}''-\boldsymbol{\theta}'))\rangle \ge W\gamma_\alpha \|\boldsymbol{\mathcal{X}}''-\boldsymbol{\mathcal{X}}'\|_F^2.$$

Proof. Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{n_1 n_2 n_3}$, we have

$$\begin{aligned}
\frac{\partial \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta})}{\partial \theta_p} &= -\sum_{t=1}^{W}\sum_{l=1}^{K} \frac{\dot{f}_l(\theta_p)}{f_l(\theta_p)} \mathbb{1}_{\{\mathrm{vec}(\boldsymbol{\mathcal{Y}}^t)_p = l\}}, \\
\frac{\partial^2 \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta})}{\partial \theta_p^2} &= \sum_{t=1}^{W}\sum_{l=1}^{K} \left(\frac{\dot{f}_l^2(\theta_p)}{f_l^2(\theta_p)} - \frac{\ddot{f}_l(\theta_p)}{f_l(\theta_p)} \right) \mathbb{1}_{\{\mathrm{vec}(\boldsymbol{\mathcal{Y}}^t)_p = l\}},
\end{aligned}$$

and $\frac{\partial^2 \mathscr{F}_{\mathscr{Y}}(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_q} = 0$ if $p \ne q$. Recalling (8) implies

$$\begin{aligned}
&\langle \boldsymbol{\theta}''-\boldsymbol{\theta}', \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathscr{F}_{\mathscr{Y}}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta}''-\boldsymbol{\theta}')\rangle \\
&= \sum_p \frac{\partial^2 \mathscr{F}_{\mathscr{Y}}(\tilde{\boldsymbol{\theta}})}{\partial \theta_p^2} (\theta_p''-\theta_p')^2 \\
&\ge W\gamma_\alpha \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} (\mathcal{X}''_{ijk} - \mathcal{X}'_{ijk})^2 \\
&= W\gamma_\alpha \|\boldsymbol{\mathcal{X}}''-\boldsymbol{\mathcal{X}}'\|_F^2.
\end{aligned}$$

□

With Lemma 1 and Lemma 2, we now prove Theorem 1.

Proof of Theorem 1. Recalling the fact that $\hat{\boldsymbol{\mathcal{X}}}, \widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{K}_{r,s}$, we obtain the first bound since

$$\begin{aligned}
&\frac{\|\hat{\boldsymbol{\mathcal{X}}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_F}{\sqrt{n_1 n_2 n_3}} \\
&\le \frac{\|\hat{\boldsymbol{\mathcal{L}}} - \widetilde{\boldsymbol{\mathcal{L}}}\|_F}{\sqrt{n_1 n_2 n_3}} + \frac{\|\hat{\boldsymbol{\mathcal{C}}} - \widetilde{\boldsymbol{\mathcal{C}}}\|_F}{\sqrt{n_1 n_2 n_3}} \\
&\le 2\alpha + \frac{2\alpha\sqrt{2s}}{\sqrt{n_1 n_2 n_3}} = 2\alpha\left(1 + \sqrt{\frac{2s}{n_1 n_2 n_3}}\right).
\end{aligned}$$

Next, we aim to obtain the second bound. We first define $\mathbb{K}_f := \{\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} : (\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}\}$. Since the objective term $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$ is continuous in any $\boldsymbol{\mathcal{X}} \in \mathbb{K}_f$ and $\mathbb{K}_f$ is a compact set, $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$ then achieves a minimum in $\mathbb{K}_f$. Suppose that $\hat{\boldsymbol{\mathcal{X}}} = \hat{\boldsymbol{\mathcal{L}}} + \hat{\boldsymbol{\mathcal{C}}} \in \mathbb{K}_f$ minimizes $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$. Combining (57) with Lemma 1 and Lemma 2 implies that

$$\begin{aligned}
F_{\mathscr{Y}}(\hat{\boldsymbol{\mathcal{X}}}) \ge& F_{\mathscr{Y}}(\widetilde{\boldsymbol{\mathcal{X}}}) - 6L_\alpha\sqrt{rWn_{(1)}n_3\log(d)}(\|\hat{\boldsymbol{\mathcal{X}}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_F \\
&+ 2\alpha\sqrt{2s}) - 4W\alpha s L_\alpha + \frac{W\gamma_\alpha}{2}\|\hat{\boldsymbol{\mathcal{X}}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_F^2
\end{aligned}$$

holds with probability at least $1-1/d$. Note that $F_{\mathscr{Y}}(\hat{\boldsymbol{\mathcal{X}}}) \le F_{\mathscr{Y}}(\widetilde{\boldsymbol{\mathcal{X}}})$. We have that

$$\begin{aligned}
&\frac{W\gamma_\alpha}{2}\|\hat{\boldsymbol{\mathcal{X}}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_F^2 \\
&\le 6L_\alpha\sqrt{rWn_{(1)}n_3\log(d)}(\|\hat{\boldsymbol{\mathcal{X}}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_F + 2\alpha\sqrt{2s}) + 4W\alpha s L_\alpha
\end{aligned}$$

holds with probability at least $1-1/d$. Recalling the fact that $u \le \max\{2a, 2b\}$ holds if $u \le a+b$, we then have

$$\frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}^*\|_F}{\sqrt{n_1 n_2 n_3}} \le U_\alpha$$

with the same probability, where

$$\begin{aligned}
U_\alpha = \max\Bigg\{ &\frac{24L_\alpha\sqrt{r\log(d)}}{\gamma_\alpha\sqrt{Wn_{(2)}}}, \\
&4\sqrt{\frac{3\alpha L_\alpha\sqrt{2rsn_{(1)}\log(d)}}{\gamma_\alpha\sqrt{W}n_1 n_2 n_3} + \frac{\alpha L_\alpha s}{\gamma_\alpha n_1 n_2 n_3}} \Bigg\},
\end{aligned}$$

which completes the proof. □

B. Proof of Theorem 2

The proof of Theorem 2 borrows ideas from [7] and [8]. Without loss of generality, we assume that $n_1 \ge n_2$ in the proof. Here, we first need some lemmas before giving the upper bound.

Lemma 3. *Define $\mathbb{K}_f := \{\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} : (\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}\}$. There exists a set $\mathbb{S} \in \mathbb{K}_f$ with*

$$|\mathbb{S}| \ge \exp\left(\frac{rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor}{16}\right) \tag{58}$$

such that for any $\gamma \in (0,1]$, the following properties hold:

(i). *For all $\boldsymbol{\mathcal{T}} \in \mathbb{S}$, $|\mathcal{T}_{ijk}| = 0$ or $\alpha\gamma$, $\forall (i,j,k) \in [n_1] \times [n_2] \times [n_3]$.*

(ii). *For all $\boldsymbol{\mathcal{T}}^{(l)}, \boldsymbol{\mathcal{T}}^{(m)} \in \mathbb{S}$, $l \ne m$,*

$$\|\boldsymbol{\mathcal{T}}^{(l)} - \boldsymbol{\mathcal{T}}^{(m)}\|_F^2 \ge \alpha^2\gamma^2\left(\frac{n_1 n_2 n_2}{2} - s\right).$$

Proof. A probabilistic argument is utilized for the proof process. We construct the set $\mathbb{S}$ by drawing $\lceil \exp(\frac{rn_1n_3 - r\lfloor\frac{s}{n_2n_3}\rfloor}{16}) \rceil$ random tensors from the following distribution. Each tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ consists of 'sub-tensors' of size $n_1 \times r \times n_3$, stacked from left to right. Consider the first 'sub-tensor', i.e., $\mathcal{T}_{ijk}$ for $i \in [n_1]$, $j \in [r]$, and $k \in [n_3]$. We fix the locations of $\lfloor\frac{s}{n_2}\rfloor$ entries in each lateral slice and set their values to be 0. The remaining $rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor$ entries of it are i.i.d symmetric random variables taking values $\pm\alpha\gamma$ with equal probability. Then, $\boldsymbol{\mathcal{T}}$ will be constructed by copying the first 'sub-tensor' as many times as will fit. More specifically, for all $i \in [n_1]$, $j \in \{r+1, \cdots, n_2\}$, and $k \in [n_3]$,

$$\mathcal{T}_{ijk} = \mathcal{T}_{ij'k}, \text{ where } j' = j(\mathrm{mod}\ r) + 1.$$

Then $\boldsymbol{\mathcal{T}}$ can be regarded as $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}}$, where $\mathrm{rank_t}(\boldsymbol{\mathcal{L}}) \le r$ and $\|\boldsymbol{\mathcal{C}}\|_0 \le s$. We can further check that

$$\|\boldsymbol{\mathcal{T}}\|_\infty = \alpha\gamma \le \alpha \text{ and } \|\boldsymbol{\mathcal{C}}\|_\infty = \alpha\gamma \le \alpha.$$

Thus we have $\mathbb{S} \subset \mathbb{K}_f$. It remains to show that $\mathbb{S}$ satisfies the second property. Note that the locations of the zero entries are the same for all tensors drawn from the above distribution. For two tensors $\boldsymbol{\mathcal{T}}$ and $\tilde{\boldsymbol{\mathcal{T}}}$ drawn as above, we have

$$\begin{aligned}
&\|\boldsymbol{\mathcal{T}} - \tilde{\boldsymbol{\mathcal{T}}}\|_F^2 \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}(\mathcal{T}_{ijk} - \tilde{\mathcal{T}}_{ijk})^2 \\
&\geq \lfloor \frac{n_2}{r} \rfloor \sum_{i=1}^{n_1}\sum_{j=1}^{r}\sum_{k=1}^{n_3}(\mathcal{T}_{ijk} - \tilde{\mathcal{T}}_{ijk})^2 \\
&= 4\alpha^2\gamma^2 \lfloor \frac{n_2}{r} \rfloor \sum_{t=1}^{q} \delta_t := 4\alpha^2\gamma^2 \lfloor \frac{n_2}{r} \rfloor Z(\boldsymbol{\mathcal{T}}, \tilde{\boldsymbol{\mathcal{T}}}),
\end{aligned}$$

where δ_t's are independent $0/1$ Bernoulli random variables with mean $1/2$ and $q := rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor$. Applying Hoeffding's inequality and a union bound demonstrates that

$$\begin{aligned}
&\mathbb{P}\left(\min_{\boldsymbol{\mathcal{T}} \neq \tilde{\boldsymbol{\mathcal{T}}} \in \mathbb{S}} Z(\boldsymbol{\mathcal{T}}, \tilde{\boldsymbol{\mathcal{T}}}) \leq \frac{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor}{4} \right) \\
&\leq \binom{|\mathbb{S}|}{2} \exp(-\frac{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor}{8}).
\end{aligned}$$

Recalling (58), the right-hand side of the above inequality is less than 1, which indicates that

$$\begin{aligned}
\|\boldsymbol{\mathcal{T}} - \tilde{\boldsymbol{\mathcal{T}}}\|_F^2 &> \alpha^2\gamma^2 \lfloor \frac{n_2}{r} \rfloor \left(rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor \right) \\
&\geq \alpha^2\gamma^2 \left(\frac{n_1n_2n_3}{2} - s \right)
\end{aligned}$$

holds for all tensors $\boldsymbol{\mathcal{T}}, \tilde{\boldsymbol{\mathcal{T}}} \in \mathbb{S}$ such that $\boldsymbol{\mathcal{T}} \neq \tilde{\boldsymbol{\mathcal{T}}}$ with a nonzero probability, where the second inequality follows from the fact that $\lfloor x \rfloor \geq x/2$ for all $x \geq 1$. Now, we have proved the second property. □

Before continuing our discussion, we introduce some definitions in information theory. For random variables $(X, Y) \sim p(x, y)$, the mutual information $I(X, Y)$ is defined as

$$I(X, Y) = -\iint p(x, y) \log\left(\frac{p(x)p(y)}{p(x, y)} \right) dxdy,$$

where $p(x)$ and $p(y)$ denote the marginal distributions of x and y, respectively. The entropy $H(X)$ and conditional entropy $H(X|Y)$ are respectively defined as

$$\begin{aligned}
H(X) &= -\int p(x) \log(x) dx, \\
H(X|Y) &= -\iint p(x, y) \log(p(x|y)) dxdy.
\end{aligned}$$

By these definitions, we have

$$I(X, Y) = H(Y) - H(Y|X). \tag{59}$$

Lemma 4. *Follow the definition of $\mathbb{S}$ in Lemma 3. Suppose $\boldsymbol{\mathcal{T}} \in \mathbb{S}$ is generated uniformly at random and $\boldsymbol{\mathcal{X}} := \boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}$ where $\boldsymbol{\mathcal{N}}$ contains i.i.d. Gaussian entries with zero mean and variance σ^2. Then we have the following result about the mutual information $I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}})$:*

$$I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}) \leq \frac{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor}{2} \log\left(1 + \left(\frac{\alpha\gamma}{\sigma} \right)^2 \right).$$

Proof. Recalling (59), we have

$$\begin{aligned}
I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}) &= H(\boldsymbol{\mathcal{X}}) - H(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{T}}) \\
&= H(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}) - H(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}|\boldsymbol{\mathcal{T}}) \\
&= H(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}) - H(\boldsymbol{\mathcal{N}}).
\end{aligned}$$

Let $\boldsymbol{\mathcal{E}}$ denote a tensor whose all entries are i.i.d. Rademacher random variables. We obtain

$$\begin{aligned}
&H(\boldsymbol{\mathcal{T}} \odot \boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{N}}) \\
&= H((\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}) \odot \boldsymbol{\mathcal{E}}) \geq H((\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}) \odot \boldsymbol{\mathcal{E}}|\boldsymbol{\mathcal{E}}) \\
&= H(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}),
\end{aligned}$$

where $\boldsymbol{\mathcal{T}} \odot \boldsymbol{\mathcal{E}}$ denotes the entry-wise product of $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{E}}$. Let $\widehat{\boldsymbol{\mathcal{T}}} = \boldsymbol{\mathcal{T}} \odot \boldsymbol{\mathcal{E}}$. Then it holds that

$$I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}) \leq H(\widehat{\boldsymbol{\mathcal{T}}} + \boldsymbol{\mathcal{N}}) - H(\boldsymbol{\mathcal{N}}).$$

For $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{T}}} + \boldsymbol{\mathcal{N}})$ of length $n_1n_2n_3$, the covariance matrix is calculated as

$$\boldsymbol{\Sigma} := \mathbb{E}[\mathrm{vec}(\widehat{\boldsymbol{\mathcal{T}}} + \boldsymbol{\mathcal{N}})\, \mathrm{vec}(\widehat{\boldsymbol{\mathcal{T}}} + \boldsymbol{\mathcal{N}})^T].$$

Applying Theorem 8.6.5 in [48] implies

$$\begin{aligned}
&H(\widehat{\boldsymbol{\mathcal{T}}} + \boldsymbol{\mathcal{N}}) \\
&\leq \frac{1}{2} \log\{(2\pi e)^{n_1n_2n_3} \det(\boldsymbol{\Sigma})\} \\
&= \frac{1}{2} \log\{(2\pi e)^{n_1n_2n_3} (\alpha^2\gamma^2 + \sigma^2)^{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor} \sigma^{2n_2\lfloor \frac{s}{n_2} \rfloor}\},
\end{aligned}$$

where the last equality follows from the fact that $\widehat{\boldsymbol{\mathcal{T}}}$ has $n_2\lfloor \frac{s}{n_2} \rfloor$ non-zero entries. On the other hand, we have

$$H(\boldsymbol{\mathcal{N}}) = \frac{1}{2} \log\{(2\pi e)^{n_1n_2n_3} \sigma^{2n_1n_2n_3}\}.$$

Therefore, we obtain that

$$\begin{aligned}
I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}) &\leq \frac{1}{2} \log\left(\frac{(\alpha^2\gamma^2 + \sigma^2)^{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor} \sigma^{2n_2\lfloor \frac{s}{n_2} \rfloor}}{\sigma^{2n_1n_2n_3}} \right) \\
&= \frac{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor}{2} \log\left(1 + \left(\frac{\alpha\gamma}{\sigma} \right)^2 \right).
\end{aligned}$$

□

Proof of Theorem 2. Choose ϵ such that

$$\epsilon^2 = \min\left\{ \frac{(1 - 2C_0)\alpha^2}{8}, \frac{C_1\sigma^2}{W} \frac{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor - 64}{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor} \right\}, \tag{60}$$

where the constant $C_1 > 0$ will be determined later. Let γ be such that

$$\frac{2\epsilon}{\alpha} \sqrt{\frac{2n_1n_2n_3}{n_1n_2n_3 - 2s}} \leq \gamma \leq \frac{2\epsilon}{\alpha} \sqrt{\frac{2}{1 - 2C_0}} \leq 1. \tag{61}$$

Consider $\boldsymbol{\mathcal{X}} \in \mathbb{S}$ drawn uniformly at random. Let $\boldsymbol{\mathcal{Y}}^i = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{N}}^i$ $(\forall i \in \{1, \cdots, W\})$. Then we have

$$\begin{aligned}
I(\boldsymbol{\mathcal{X}}, \mathscr{Y}) &\leq \sum_{i=1}^{W} I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}^i) \\
&\leq \frac{W(n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor)}{2} \log\left(1 + \left(\frac{\alpha\gamma}{\sigma} \right)^2 \right),
\end{aligned} \tag{62}$$

where the last inequality follows from Lemma 4.

For the sake of contradiction, we suppose that there exists an algorithm to return an estimate $\hat{\boldsymbol{\mathcal{X}}}$ for any $\boldsymbol{\mathcal{X}} \in \mathbb{S}$ and corresponding measurements $\boldsymbol{\mathcal{Y}}$, so that

$$\frac{\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F^2}{n_1 n_2 n_3} \leq \epsilon^2 \tag{63}$$

holds with probability at least $1/4$. Define

$$\boldsymbol{\mathcal{X}}^* := \arg\min_{\boldsymbol{\mathcal{T}} \in \mathbb{S}} \|\boldsymbol{\mathcal{T}} - \hat{\boldsymbol{\mathcal{X}}}\|_F^2. \tag{64}$$

We claim that if (63) holds, then $\boldsymbol{\mathcal{X}}^* = \boldsymbol{\mathcal{X}}$. To see this, for any $\boldsymbol{\mathcal{T}} \in \mathbb{S}$ with $\boldsymbol{\mathcal{T}} \neq \boldsymbol{\mathcal{X}}$, using the second property of Lemma 3 and (61) implies

$$\|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{X}}\|_F > \alpha\gamma\sqrt{n_1 n_2 n_3/2 - s} \geq 2\sqrt{n_1 n_2 n_3}\epsilon. \tag{65}$$

Combining (64) and (65) deduces

$$\begin{aligned}\|\boldsymbol{\mathcal{T}} - \hat{\boldsymbol{\mathcal{X}}}\|_F &\geq \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{X}}\|_F - \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F \\ &> 2\sqrt{n_1 n_2 n_3}\epsilon - \sqrt{n_1 n_2 n_3}\epsilon = \sqrt{n_1 n_2 n_3}\epsilon.\end{aligned}$$

Since $\boldsymbol{\mathcal{X}} \in \mathbb{S}$ is an candidate for $\boldsymbol{\mathcal{X}}^*$, we thus have

$$\|\boldsymbol{\mathcal{X}}^* - \hat{\boldsymbol{\mathcal{X}}}\|_F \leq \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F \leq \sqrt{n_1 n_2 n_3}\epsilon.$$

Therefore, if (63) holds, then $\|\boldsymbol{\mathcal{X}}^* - \hat{\boldsymbol{\mathcal{X}}}\|_F < \|\boldsymbol{\mathcal{T}} - \hat{\boldsymbol{\mathcal{X}}}\|_F$ for any $\boldsymbol{\mathcal{T}} \in \mathbb{S}$ with $\boldsymbol{\mathcal{T}} \neq \boldsymbol{\mathcal{X}}$, and hence we must have $\boldsymbol{\mathcal{X}}^* = \boldsymbol{\mathcal{X}}$. By the assumption that (63) holds with probability at least $\frac{1}{4}$, we thus have

$$\mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) \leq \frac{3}{4}. \tag{66}$$

On the other hand, by Fano's inequality,

$$\begin{aligned}\mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) &\geq \frac{H(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) - 1}{\log(|\mathbb{S}|)} = \frac{H(\boldsymbol{\mathcal{X}}) - I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) - 1}{\log(|\mathbb{S}|)} \\ &\geq 1 - \frac{I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) + 1}{\log(|\mathbb{S}|)}\end{aligned}$$

Plugging in $|\mathbb{S}|$ from (58) and $I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ from (62), and noting the inequality $\log(1+u) \leq u$, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) \geq & 1 - \frac{16}{rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor} \\ & \cdot \left(\frac{n_1n_2n_3 - n_2\lfloor\frac{s}{n_2}\rfloor}{2} W \left(\frac{\alpha\gamma}{\sigma}\right)^2 + 1\right).\end{aligned}$$

Combining this with (61) and (66) implies

$$\begin{aligned}\frac{1}{4} \leq & \frac{16}{rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor} \\ & \cdot \left(\frac{4W}{1-2C_0}\left(n_1n_2n_3 - n_2\lfloor\frac{s}{n_2}\rfloor\right)\left(\frac{\epsilon}{\sigma}\right)^2 + 1\right),\end{aligned}$$

which shows that

$$\epsilon^2 \geq \frac{(1-2C_0)\sigma^2}{256W} \frac{rn_1n_3 - r\lfloor\frac{s}{n_2}\rfloor - 64}{n_1n_2n_3 - n_2\lfloor\frac{s}{n_2}\rfloor}.$$

Setting $C_1 < \frac{1-2C_0}{256}$ in (60) will lead to a contradiction. Hence (63) must fail to hold with probability at least $3/4$, which completes the proof. □

C. Proof of Theorem 3

We first require some definitions in information theory given below.

Definition 6. *For any integer $K > 0$, a function $f : \mathbb{R} \to \mathscr{S}_K$ is called a K-link function, where $\mathscr{S}_K$ is the K-dimensional probability simplex.*

Definition 7. *For a K-link function f and tensors $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the squared Hellinger distance is defined as*

$$\mathrm{d_H^2}(f(\boldsymbol{\mathcal{A}}), f(\boldsymbol{\mathcal{B}})) := \frac{1}{n_1n_2n_3} \sum_{i,j,k} \sum_{l=1}^{K} \mathrm{d_H^2}(f_l(\mathcal{A}_{ijk}), f_l(\mathcal{B}_{ijk})), \tag{67}$$

and the Kullback-Leibler divergence is

$$\mathrm{KL}(f(\boldsymbol{\mathcal{A}}), f(\boldsymbol{\mathcal{B}})) := \frac{1}{n_in_2n_3} \sum_{i,j,k} \sum_{l=1}^{K} f_l(\mathcal{A}_{ijk}) \log\left(\frac{f_l(\mathcal{A}_{ijk})}{f_l(\mathcal{B}_{ijk})}\right),$$

where $\mathrm{d_H^2}(f_l(a), f_l(b)) := \sum_{l=1}^{K}(\sqrt{f_l(a)} - \sqrt{f_l(b)})^2$ and f_l denotes the l-th dimensional function of f.

The following lemma indicates that the Euclidean distance is upper bounded by the squared Hellinger distance and the the Kullback-Leibler divergence.

Lemma 5. *For any $p > 0$ and a p-link function f and any $\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ satisfying $\|\boldsymbol{\mathcal{X}}\|_\infty \leq 2\alpha$ and $\|\hat{\boldsymbol{\mathcal{X}}}\|_\infty \leq 2\alpha$, we have*

$$\begin{aligned}\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F^2 &\leq \frac{n_1n_2n_3}{K_\alpha} \mathrm{d_H}^2(f(\boldsymbol{\mathcal{X}}), f(\hat{\boldsymbol{\mathcal{X}}})) \\ &\leq \frac{n_1n_2n_3}{K_\alpha} \mathrm{KL}(f(\boldsymbol{\mathcal{X}}), f(\hat{\boldsymbol{\mathcal{X}}})),\end{aligned}$$

where K_α is defined as in (29)*.*

Proof. The first inequality follows from the definition of K_α in (29) . The proof of the second inequality borrows the ideas from [49, Lemma 2.4]. Since $-\log(x+1) \geq -x$ for $x > -1$ and $\sum_{l=1}^{p} f_l(x) = 1$, we have

$$\begin{aligned}&\mathrm{KL}(f(\boldsymbol{\mathcal{X}}), f(\hat{\boldsymbol{\mathcal{X}}})) \\ &= \frac{1}{n_1n_2n_3} \sum_{i,j,k} \sum_{l=1}^{p} f_l(\mathcal{X}_{ijk}) \log\left(\frac{f_l(\mathcal{X}_{ijk})}{f_l(\hat{\mathcal{X}}_{ijk})}\right) \\ &= \frac{2}{n_1n_2n_3} \sum_{i,j,k} \sum_{l=1}^{p} f_l(\mathcal{X}_{ijk}) \log\left(\sqrt{\frac{f_l(\mathcal{X}_{ijk})}{f_l(\hat{\mathcal{X}}_{ijk})}}\right) \\ &= -\frac{2}{n_1n_2n_3} \sum_{i,j,k} \sum_{l=1}^{p} f_l(\mathcal{X}_{ijk}) \log\left(\left[\sqrt{\frac{f_l(\hat{\mathcal{X}}_{ijk})}{f_l(\mathcal{X}_{ijk})}} - 1\right] + 1\right) \\ &\geq -\frac{2}{n_1n_2n_3} \sum_{i,j,k} \sum_{l=1}^{p} f_l(\mathcal{X}_{ijk}) \left[\sqrt{\frac{f_l(\hat{\mathcal{X}}_{ijk})}{f_l(\mathcal{X}_{ijk})}} - 1\right] \\ &= -\frac{2}{n_1n_2n_3} \sum_{i,j,k} \left(\sum_{l=1}^{p} \sqrt{f_l(\hat{\mathcal{X}}_{ijk}) f_l(\mathcal{X}_{ijk})} - 1\right) \\ &= \mathrm{d_H^2}(f(\boldsymbol{\mathcal{X}}), f(\hat{\boldsymbol{\mathcal{X}}})).\end{aligned} \tag{68}$$

□

Proposition 5 (See [50])**.** *Consider a finite sequence of independent random matrices $\{\boldsymbol{Z}_i\}_{i=1}^n \in \mathbb{R}^{p\times q}$ that satisfy $\mathbb{E}[\boldsymbol{Z}_i] = \boldsymbol{0}$ and for some $U > 0$, $\|\boldsymbol{Z}_i\| \le U$ for all $i \in [n]$. Let $h \ge 1$ and assume $n \ge n^* := U^2\log(d)/(9\sigma_Z^2)$. Then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\right\|^h\right] \le \left(\frac{2ehc^{*2}\sigma_Z^2\log(d)}{n}\right)^{h/2},$$

where $d = p+q$, $c^* = 1+\sqrt{3}$ and

$$\sigma_Z^2 = \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\boldsymbol{Z}_i\boldsymbol{Z}_i^T]\right\|, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\boldsymbol{Z}_i^T\boldsymbol{Z}_i]\right\|\right\}.$$

In our proof of Theorem 3, it is convenient to consider the function

$$\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) := F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - F_{\mathscr{Y}}(\boldsymbol{0})$$

instead of considering $F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$ itself. With this definition, we have the following result.

Lemma 6. *Let $\mathbb{K}_{r,s}^* \subset \mathbb{R}^{n_1\times n_2\times n_3}$ be defined as in* (20) *and $\mathbb{K}_f^* := \{\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}} : (\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{C}}) \in \mathbb{K}_{r,s}^*\}$. We assume that $n_{(1)} \ge \log(d)/9$ where $d = (n_1+n_2)n_3$. Then with probability at most $1/d$, it holds that*

$$\begin{aligned}&\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]| \\ &\ge C_0\alpha L_\alpha\left(n_3n_{(1)}\log(d)\sqrt{Wn_{(2)}r} + s\sqrt{Wn_3n_{(1)}\log(d)}\right)\end{aligned}$$

where C_0 is an absolute constant and the expectation is over the draw of $\mathscr{Y}$.

Proof. Applying Markov's inequality implies that, for any $h > 0$, the following holds

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]| \ge C_0\alpha L_\alpha \right.\\ &\qquad \left.\cdot(n_3n_{(1)}\log(d)\sqrt{Wn_{(2)}r} + s\sqrt{Wn_3n_{(1)}\log(d)})\right)\\ &= \mathbb{P}\left(\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]|^h \ge (C_0\alpha L_\alpha \right.\\ &\qquad \left.\cdot(n_3n_{(1)}\log(d)\sqrt{Wn_{(2)}r} + s\sqrt{Wn_3n_{(1)}\log(d)}))^h\right)\\ &\le \frac{\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]|^h\right]}{(C_0\alpha L_\alpha\sqrt{W}H)^h},\end{aligned} \tag{69}$$

where $H := n_3n_{(1)}\log(d)\sqrt{n_{(2)}r} + s\sqrt{n_3n_{(1)}\log(d)}$.

Since we can rewrite the definition of $\bar{F}_{\mathscr{Y}}$ as

$$\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) = \sum_{i,j,k}\left(\sum_{t=1}^W\sum_{l=1}^K \mathbb{1}_{[\mathcal{Y}_{ijk}^t=1]}\log\left(\frac{f_l(\mathcal{X}_{ijk})}{f_l(0)}\right)\right),$$

by a symmetrization argument [51, Lemma 6.3], we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} \left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\right|^h\right]\\ &\le 2^h\,\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*}\left|\sum_{i,j,k}\epsilon_{ijk}\left(\sum_{t=1}^W\sum_{l=1}^K \mathbb{1}_{[\mathcal{Y}_{ijk}^t=1]}\log\left(\frac{f_l(\mathcal{X}_{ijk})}{f_l(0)}\right)\right)\right|^h\right],\end{aligned}$$

where ϵ_{ijk}'s are i.i.d. Rademacher random variables and the expectation in the upper bound is with respect to $\mathscr{Y}$ as well as the ϵ_{ijk}'s. Now we leverage a contraction principle [51, Theorem 4.12] to bound the latter term. By the definition of L_α in (9) and the assumption that $\|\boldsymbol{\mathcal{X}}\|_\infty \le 2\alpha$, we have that

$$\frac{1}{L_\alpha}\log\left(\frac{f_l(\mathcal{X}_{ijk})}{f_l(0)}\right),\ \forall\ l \in [K]$$

are all contractions that vanish at zero for all $(i,j,k) \in [n_1]\times[n_2]\times[n_3]$. Thus, up to a factor 2, the expected value of the supremum can only increase when the above terms are replaced by $\mathcal{X}_{ijk}$, respectively, i.e.,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} \left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}) - \mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\right|^h\right]\\ &\le 2^h(2L_\alpha)^h\,\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*}\left|\sum_{i,j,k}\epsilon_{ijk}\left(\sum_{t=1}^W\sum_{l=1}^K \mathbb{1}_{[\mathcal{Y}_{ijk}^t=l]}\mathcal{X}_{ijk}\right)\right|^h\right]\\ &= (4L_\alpha)^h\,\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\langle\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}\rangle|^h\right],\end{aligned} \tag{70}$$

where $\boldsymbol{\mathcal{E}}$ denotes the Rademacher tensor with entries being given by ϵ_{ijk}, $\boldsymbol{\mathcal{M}}$ denotes the tensor whose all entries are valued in the set $\{1,\cdots,W\}$, and $\odot$ denotes Hadamard product. Note that $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}}$, $\|\boldsymbol{\mathcal{L}}\|_* \le \alpha\sqrt{n_1n_2n_3r}$ and $\|\boldsymbol{\mathcal{C}}\|_1 \le \alpha s$. We have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}_f^*} |\langle\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}\rangle|^h\right]\\ &\overset{(a)}{\le} \mathbb{E}\left[\sup_{\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}\in\mathbb{K}_f^*} (|\langle\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{L}}\rangle| + |\langle\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{C}}\rangle|)^h\right]\\ &\overset{(b)}{\le} \mathbb{E}\left[\sup_{\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}\in\mathbb{K}_f^*} (|\langle\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{L}}\rangle| + \|\boldsymbol{\mathcal{M}}\|_\infty\alpha s)^h\right]\\ &\overset{(c)}{\le} \mathbb{E}\left[\left(\sup_{\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}\in\mathbb{K}_f^*} \|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\|\|\boldsymbol{\mathcal{L}}\|_* + \|\boldsymbol{\mathcal{M}}\|_\infty\alpha s\right)^h\right]\\ &\overset{(d)}{\le} \mathbb{E}\left[(\|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\|\alpha(\sqrt{n_1n_2n_3r}+s))^h\right]\\ &= (\alpha(\sqrt{n_1n_2n_3r}+s))^h\,\mathbb{E}\left[\|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\|^h\right],\end{aligned} \tag{71}$$

where (a) holds from the triangle inequality. (b) follows from the Cauchy-Schwarz inequality and the fact that $\|\boldsymbol{\mathcal{C}}\|_1 \le \alpha s$. (c) holds from the fact that tensor spectral norm $\|\cdot\|$ is dual to the tensor nuclear norm $\|\cdot\|_*$. (d) follows from the assumption that $\boldsymbol{\mathcal{L}} \le \alpha\sqrt{n_1n_2n_3r}$ and the fact that $\|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\| = \|\mathrm{bcirc}(\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}})\| \ge \|\boldsymbol{\mathcal{M}}\|_\infty$ where the inequality is obvious by the definition of matrix SVD and matrix spectral norm. Next, we aim to bound $\mathbb{E}\left[\|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\|^h\right]$. Consider

$$\mathcal{Z}_{ijk} = W^{-1}[\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}]_{ijk} = W^{-1}\epsilon_{ijk}\sum_{t=1}^W\sum_{l=1}^K \mathbb{1}_{\{\mathcal{Y}_{ijk}^t=l\}},$$

and let $\boldsymbol{\mathcal{Z}}_{ijk}$ denote the tensor whose (i,j,k)-th entry is $\mathcal{Z}_{ijk}$ and other entries are all zero. Then we need to bound

$$\mathbb{E}\left[\|\boldsymbol{\mathcal{E}}\odot\boldsymbol{\mathcal{M}}\|^h\right]=W^h\,\mathbb{E}\left[\left\|\sum_{i,j,k}\boldsymbol{\mathcal{Z}}_{ijk}\right\|^h\right]. \tag{72}$$

Recalling $\|\sum_{i,j,k}\boldsymbol{\mathcal{Z}}_{ijk}\|=\|\sum_{i,j,k}\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})\|$, we turn to bound $E[\|\sum_{i,j,k}\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})\|^h]$. It is not hard to check that $\mathbb{E}(\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk}))=\mathbf{0}$ and $\|\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})\|\le 1$. By a similar discussion to the proof of Lemma 1, we obtain

$$\left\|\frac{1}{n_1n_2n_3}\sum_{i,j,k}\mathbb{E}[\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T]\right\|\le\frac{1}{Wn_1},$$

and

$$\left\|\frac{1}{n_1n_2n_3}\sum_{i,j,k}\mathbb{E}[\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})^T\mathrm{bcirc}(\boldsymbol{\mathcal{Z}}_{ijk})]\right\|\le\frac{1}{Wn_2}.$$

Recalling $n_{(1)}\ge\log(d)/9$, applying Proposition 5 with $n:=n_1n_2n_3$ that obviously satisfies $n\ge n_3n_{(2)}\log(d)/9$, $U=1$, $\sigma_Z^2=1/(Wn_{(2)})$, $d=(n_1+n_2)n_3$, and $t=\log(d)$ implies that

$$\mathbb{E}\left[\left\|\sum_{i,j,k}\boldsymbol{\mathcal{Z}}_{ijk}\right\|^h\right]\le(2ehc^*n_3n_{(1)}\log(d)/W)^{h/2}$$

Combining this inequality with (70), (71), and (72) implies that

$$\begin{aligned}&\left(\mathbb{E}\left[\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}^*_{r,s}}\left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})-\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\right|^h\right]\right)^{\frac{1}{h}}\\&\le 4\sqrt{2(1+\sqrt{3})e}\alpha L_\alpha\\&\quad\cdot\left(n_3n_{(1)}\sqrt{Whn_{(2)}r\log(d)}+s\sqrt{Wn_3n_{(1)}\log(d)}\right).\end{aligned} \tag{73}$$

Let $h=\log[(n_1+n_2)n_3]$ and plugging (73) into (69) show that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\mathcal{X}}\in\mathbb{K}^*_{r,s}}\left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})-\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\right|\ge C_0\alpha L_\alpha\right.\\&\qquad\left.\cdot(n_3n_{(1)}\log(d)\sqrt{Wn_{(2)}r}+s\sqrt{Wn_3n_{(1)}\log(d)})\right)\end{aligned}$$

is upper bounded by

$$\left(\frac{4\sqrt{2(1+\sqrt{3})e}}{C_0}\right)^{\log[(n_1+n_2)n_3]}\le\frac{1}{(n_1+n_2)n_3},$$

provided $C_0\ge 4\sqrt{2(1+\sqrt{3})}e^{3/2}$, which completes the proof. □

Armed with Proposition 5 and Lemma 6, we can bound the square Hellinger error.

Lemma 7. *Define the K-link function $f=(f_1,f_2,\cdots,f_K)$, where f_l ($l\in[K]$) is given via* (5)*. Let $\mathbb{K}^*_{r,s}\subset\mathbb{R}^{n_1\times n_2\times n_3}$ be defined via* (20) *and $\mathbb{K}^*_f:=\left\{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}:(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}})\in\mathbb{K}^*_{r,s}\right\}$. Assume that $\boldsymbol{\mathcal{X}}\in\mathbb{K}^*_f$ and $d=(n_1+n_2)n_3$. Consider the observations $\mathscr{Y}=[\boldsymbol{\mathcal{Y}}^1,\cdots,\boldsymbol{\mathcal{Y}}^W]$ generated as in* (15) *and L_α defined as in* (9)*. Let $(\hat{\boldsymbol{\mathcal{L}}},\hat{\boldsymbol{\mathcal{C}}})$ be the solution to* (21) *and $\hat{\boldsymbol{\mathcal{X}}}=\hat{\boldsymbol{\mathcal{L}}}+\hat{\boldsymbol{\mathcal{C}}}$. Then with probability at least $1-1/d$, we have*

$$\begin{aligned}&\mathrm{d}^2_{\mathrm{H}}(f(\hat{\boldsymbol{\mathcal{X}}}),f(\boldsymbol{\mathcal{X}}))\\&\le 2C_0\alpha L_\alpha\left(\sqrt{\frac{r\log^2(d)}{n_{(2)}W}}+\frac{s\sqrt{\log(d)}}{n_{(2)}\sqrt{Wn_3n_{(1)}}}\right),\end{aligned}$$

where C_0 is an absolute constant and $d=(n_1+n_2)n_3$.

Proof. Note that for any tensor $\boldsymbol{\mathcal{T}}\in\mathbb{R}^{n_1\times n_2\times n_3}$, we have

$$\begin{aligned}&\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\\&=\mathbb{E}[F_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]\\&=W\sum_{i,j,k}\sum_{l=1}^{K}\left(f_l(\mathcal{X}_{ijk})\log\left(\frac{f_l(\mathcal{T}_{ijk})}{f_l(\mathcal{X}_{ijk})}\right)\right)\\&=-n_1n_2n_3W\mathrm{KL}(f(\boldsymbol{\mathcal{X}}),f(\boldsymbol{\mathcal{T}})),\end{aligned}$$

where the expectation is over $\mathscr{Y}$. Therefore, it holds that for any $\boldsymbol{\mathcal{T}}\in\mathbb{K}^*_f$,

$$\begin{aligned}&\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})\\&=\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]+(\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})])\\&\quad-(\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})-\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})])\\&\le\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})]+2\sup_{\boldsymbol{\mathcal{T}}\in\mathbb{K}^*_f}\left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\mathbb{E}[\bar{L}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})]\right|\\&=-n_1n_2n_3W\mathrm{KL}(f(\boldsymbol{\mathcal{X}}),f(\boldsymbol{\mathcal{T}}))\\&\quad+2\sup_{\boldsymbol{\mathcal{T}}\in\mathbb{K}^*_f}\left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\mathbb{E}[\bar{F}_{\Omega,\mathscr{Y}}(\boldsymbol{\mathcal{T}})]\right|.\end{aligned}$$

Since $\hat{\boldsymbol{\mathcal{X}}}\in\mathbb{K}^*_f$ by assumption and $\bar{F}_{\mathscr{Y}}(\hat{\boldsymbol{\mathcal{X}}})\ge\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})$, we have

$$\mathrm{KL}(f(\boldsymbol{\mathcal{X}}),f(\hat{\boldsymbol{\mathcal{X}}}))\le\frac{2}{n_1n_2n_3W}\sup_{\boldsymbol{\mathcal{T}}\in\mathbb{K}^*_f}\left|\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})-\mathbb{E}[\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{T}})]\right|.$$

Applying Lemma 6, with probability at least $1-1/d$, it holds that

$$\begin{aligned}&\mathrm{KL}(f(\boldsymbol{\mathcal{X}}),f(\hat{\boldsymbol{\mathcal{X}}}))\\&\le 2C_0\alpha L_\alpha\left(\sqrt{\frac{r\log^2(d)}{n_{(2)}W}}+\frac{s\sqrt{\log(d)}}{n_{(2)}\sqrt{Wn_3n_{(1)}}}\right).\end{aligned}$$

Finally, using Lemma 5 completes the proof. □

Proof of Theorem 3. Define $\bar{F}_{\mathscr{Y}}(\boldsymbol{\mathcal{X}}):=F_{\mathscr{Y}}(\boldsymbol{\mathcal{X}})-F_{\mathscr{Y}}(\mathbf{0})$. The proof of Theorem 3 follows immediately from Lemma 7 combined with Lemma 5. □

D. Proof of Theorem 4

Before starting the proof process, we need the following Lemma 8 and Lemma 9 for preparation.

Lemma 8. *Define $\mathbb{K}_f:=\left\{\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{C}}:(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{C}})\in\mathbb{K}^*_{r,s}\right\}$. Let $\gamma\le 1$ be such that r/γ^2 is an integer, and suppose that $r/\gamma^2\le n_2$. There exists a set $\mathbb{S}\in\mathbb{K}^*_f$ with*

$$|\mathbb{S}|\ge\exp\left(\frac{rn_1n_3-r\lfloor\frac{r}{n_2}\rfloor}{16\gamma^2}\right) \tag{74}$$

satisfing the following properties:

(i). *For all* $\boldsymbol{\mathcal{T}} \in \mathbb{S}$, $|\mathcal{T}_{ijk}| = 0$ *or* $\alpha\gamma$, $\forall (i,j,k) \in [n_1] \times [n_2] \times [n_3]$.

(ii). *For all* $\boldsymbol{\mathcal{T}}^{(l)}, \boldsymbol{\mathcal{T}}^{(m)} \in \mathbb{S}$, $l \neq m$,

$$\|\boldsymbol{\mathcal{T}}^{(l)} - \boldsymbol{\mathcal{T}}^{(m)}\|_F^2 \geq \alpha^2\gamma^2 \left(\frac{n_1 n_2 n_2}{2} - s\right).$$

Proof. We also utilize a probabilistic argument for the proof process. The set $\mathbb{S}$ will be constructed by drawing

$$|\mathbb{S}| \geq \exp\left(\frac{rn_1n_3 - r\lfloor \frac{r}{n_2} \rfloor}{16\gamma^2}\right)$$

tensors $\boldsymbol{\mathcal{T}}$ independently from the following distribution. First, define $q := r/\gamma^2$ and consider a group of tensors that consists of tensor of size $n_1 \times q \times n_3$, stacked from the left to the right. We fix the locations of $\lfloor \frac{s}{n_2} \rfloor$ in each lateral slice and set their values to be 0. The remaining $qn_1n_3 - q\lfloor \frac{s}{n_2} \rfloor$ entries of it are i.i.d. symmetric random variables taking values $\pm\alpha\gamma$ with equal probability. Then, the tensor $\boldsymbol{\mathcal{T}}$ will be generated by copying the first tensor of size $n_1 \times q \times n_3$ as many times as will fit. More specifically, for all $i \in [n_1]$, $j \in \{r+1, \cdots, n_2\}$, and $k \in [n_3]$,

$$\mathcal{T}_{ijk} = \mathcal{T}_{ij'k}, \text{ where } j' = j (\text{mod } r) + 1.$$

Therefore, $\boldsymbol{\mathcal{T}}$ can be regarded as $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{C}}$, where $\text{rank}_t(\boldsymbol{\mathcal{T}}) \leq q$ and $\|\boldsymbol{\mathcal{C}}\|_0 \leq s$. Furthermore, $\text{rank}_t(\boldsymbol{\mathcal{T}}) \leq q$ implies that

$$\|\boldsymbol{\mathcal{L}}\|_{\circledast} \leq \sqrt{q}\|\boldsymbol{\mathcal{L}}\|_F \leq \sqrt{r/\gamma^2}\sqrt{n_1n_2n_3}\alpha\gamma = \alpha\sqrt{n_1n_2n_3r},$$

and $\|\boldsymbol{\mathcal{C}}\|_0 \leq s$ combining with $\|\boldsymbol{\mathcal{C}}\|_\infty \leq \alpha\gamma \leq \alpha$ indicates

$$\|\boldsymbol{\mathcal{C}}\|_1 \leq \alpha s,$$

which leads to $\boldsymbol{\mathcal{T}} \in \mathbb{K}_f^*$. Thus we have shown that $\mathbb{S} \in \mathbb{K}_f^*$. It remains to verify that $\mathbb{S}$ also satisfies the property (ii), whose proof follows a similar discussion to the proof process of Lemma 3. We omit it here. □

Lemma 9. *Follow the definition of* $\mathbb{S}$ *in Lemma 8. Suppose* $\boldsymbol{\mathcal{T}} \in \mathbb{S}$ *is generated uniformly at random and* $\boldsymbol{\mathcal{X}} := \boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{N}}$ *where* $\boldsymbol{\mathcal{N}}$ *contains i.i.d. Gaussian entries with zero mean and variance* σ^2*. Then we have the following result about the mutual information* $I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}})$*:*

$$I(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}) \leq \frac{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor}{2} \log\left(1 + \left(\frac{\alpha\gamma}{\sigma}\right)^2\right).$$

Proof. Using essentially the same argument as in the proof of Lemma 4 completes the proof. □

Proof of Theorem 4. Choose ϵ such that

$$\epsilon^2 = \min\left\{\frac{1}{16}, C_2\alpha\sigma\sqrt{\frac{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor}{W(n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor)}}\right\}, \quad (75)$$

where the constant $C_2 > 0$ will be determined later. Define γ such that r/γ^2 is an integer and

$$\frac{2\sqrt{2}\epsilon}{\alpha} \leq \gamma \leq \frac{4\epsilon}{\alpha}. \quad (76)$$

This is possible since $\epsilon \leq \frac{1}{4}$ and $r, \alpha \geq 1$. As in the proof of Theorem 2, we will consider running such an algorithm on a random tensor in the set $\mathbb{S} \subset \mathbb{K}_f^*$ defined in Lemma 8. Let $\boldsymbol{\mathcal{Y}}^i = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{N}}^i$ $(\forall i \in \{1, \cdots, W\})$. Then we have

$$\begin{aligned} I(\boldsymbol{\mathcal{X}}, \mathscr{Y}) &\leq \sum_{i=1}^{W} I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}^i) \\ &\leq \frac{W(n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor)}{2} \log\left(1 + \left(\frac{\alpha\gamma}{\sigma}\right)^2\right), \end{aligned} \quad (77)$$

where the last inequality follows from Lemma 9.

We now proceed by using the same argument as in the proof of Theorem 2. Specifically, we suppose for the sake of contradiction that there exists an algorithm to return an estimate $\hat{\boldsymbol{\mathcal{X}}}$ for any $\boldsymbol{\mathcal{X}} \in \mathbb{S}$ and corresponding measurements $\boldsymbol{\mathcal{Y}}$, so that

$$\frac{\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F^2}{n_1n_2n_3} \leq \epsilon^2 \quad (78)$$

holds with probability at least $1/4$. As before, if we set

$$\boldsymbol{\mathcal{X}}^* := \arg\min_{\boldsymbol{\mathcal{T}} \in \mathbb{S}} \|\boldsymbol{\mathcal{T}} - \hat{\boldsymbol{\mathcal{X}}}\|_F^2,$$

then we can show that (78) implies $\boldsymbol{\mathcal{X}}^* = \boldsymbol{\mathcal{X}}$. Thus if (78) holds with probability at least $1/4$, then

$$\mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) \leq \frac{3}{4}. \quad (79)$$

However, by Fano's inequality, the probability that $\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}$ is at least

$$\begin{aligned} \mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) &\geq \frac{H(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) - 1}{\log(|\mathbb{S}|)} = \frac{H(\boldsymbol{\mathcal{X}}) - I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) - 1}{\log(|\mathbb{S}|)} \\ &\geq 1 - \frac{I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) + 1}{\log(|\mathbb{S}|)} \end{aligned}$$

Plugging in $|\mathbb{S}|$ from (74), $I(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ from Lemma 9, and the inequality $\log(1+u) \leq u$, indicates that

$$\begin{aligned} &\mathbb{P}(\boldsymbol{\mathcal{X}}^* \neq \boldsymbol{\mathcal{X}}) \\ &\geq 1 - \frac{16\gamma^2}{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor}\left(\frac{n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor}{2} W\left(\frac{\alpha\gamma}{\sigma}\right)^2 + 1\right). \end{aligned}$$

Combining this with (79) and the assumption (76), we obtain

$$\begin{aligned} \frac{1}{4} \leq &\frac{256\epsilon^2}{(rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor)\alpha^2} \\ &\cdot \left(8W(n_1n_2n_3 - n_2\lfloor s/n_2 \rfloor)\left(\frac{\epsilon}{\sigma}\right)^2 + 1\right). \end{aligned}$$

We now argue, as before, that the result leads to a contradiction. Specifically, if $8W(n_1n_2n_3 - n_2\lfloor s/n_2 \rfloor)(\epsilon/\sigma)^2 \leq 1$, then together with (75) this implies that

$$\alpha^2(rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor) \leq 128.$$

If we set $C_0 > 128$ in the assumption (34), then this would result in a contradiction. Thus, suppose now that $8W(n_1n_2n_3 - n_2\lfloor s/n_2 \rfloor)(\epsilon/\sigma)^2 > 1$, in which case we have

$$\epsilon^2 > \frac{\alpha\sigma}{128}\sqrt{\frac{rn_1n_3 - r\lfloor \frac{s}{n_2} \rfloor}{W(n_1n_2n_3 - n_2\lfloor \frac{s}{n_2} \rfloor)}}.$$

Then, setting $C_2 \leq \frac{1}{128}$ in (75) also leads to a contradiction, and hence (78) must fail to hold with probability at least $3/4$, which finishes the proof. □

REFERENCES

[1] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[2] A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk, "SpaRCS: Recovering low-rank and sparse matrices from compressive measurements," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1089–1097.

[3] K. Min, Z. D. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in: *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, 2010, pp. 269–278.

[4] O. Solomon, R. Cohen, Y. Zhang, Y. Yang, Q. He, J. W. Luo, R. J. V. Sloun, and Y. C. Eldar, "Deep unfolded robust PCA with application to clutter suppression in ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1051–1063, 2020.

[5] A. Reinhardt, F. Englert, and D. Christin, "Enhancing user privacy by preprocessing distributed smart meter data," in *Proc. Sustain. Internet ICT Sustain.*, 2013, pp. 1–7.

[6] A. Soni, S. Jain, J. Haupt, and S. Gonella, "Noisy matrix completion under sparse factor models," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3636–3661, 2016.

[7] P. Z. Gao, R. Wang, M. Wang, and J. H. Chow, "Low-rank matrix recovery from noisy, quantized, and erroneous measurements," *IEEE Trans. on Signal Process.*, vol. 66, no. 11, pp. 2918–2932, 2018.

[8] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[9] A. S. Lan, C. Studer, and R.G. Baraniuk, "Matrix recovery from quantized and corrupted measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4973–4977.

[10] S. A. Bhaskar, "Probabilistic low-rank matrix recovery from quantized measurements: Application to image denoising," in *Proc. Asilomar Conf. Signals, Syst. and Comput.*, 2015, pp. 541–545.

[11] S. A. Bhaskar, "Localization from connectivity: A 1-bit maximum likelihood approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2939–2953, 2016.

[12] S. A. Bhaskar, "Probabilistic low-rank matrix completion from quantized measurements," *J. Mach. Learn. Res.*, vol. 17, no. 60, pp. 1–34, 2016.

[13] Y. Cao and Y. Xie, "Categorical matrix completion," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sens. Adaptive Process.*, 2015, pp. 369–372.

[14] O. Klopp, J. Lafond, É. Moulines, and J. Salmon, "Adaptive multinomial matrix completion," *Electron. J. Statist.*, vol. 9, no. 2, pp. 2950–2975, 2015.

[15] J. Lafond, O. Klopp, É. Moulines, and J. Salmon, "Probabilistic low-rank matrix completion on finite alphabets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 1727–1735, 2014.

[16] T. Cai and W. X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3619–3647, 2013.

[17] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1959–2008, 2014.

[18] Y. Peng, D. Y. Meng, Z. B. Xu, C. Q. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vision Pattern Recog*, 2014, pp. 2949–2956.

[19] Y. Wang, J. J. Peng, Q. Zhao, Y. Leung, X. L. Zhao, and D. Y. Meng, "Hyperspectral image restoration via total variation regularized low-rank tensor decomposition," *IEEE J. Sel. Top. Appl. Earth. Observ. Remote Sens*, vol. 11, no. 4, pp. 1227–1243, 2018.

[20] F. Zhang, J. J. Wang, W. D. Wang, and C. Xu, "Low-tubal-rank plus sparse tensor recovery with prior subspace information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3492–3507, 2021.

[21] L. Baltrunas, M. Kaminskas, B. Ludwig, *et al.* "InCarMusic: context-aware music recommendations in a car," *E-Commerce and Web Technologies*, Springer, Berlin, Heidelberg, 2011, pp. 89–100.

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

[23] S. Friedland, "Variation of tensor powers and spectrat," *Linear Multilinear Algebra*, vol. 12, no. 2, pp. 81–98, 1982.

[24] L. H. Lim and P. Comon, "Multiarray signal processing: tensor decomposition meets compressed sensing," *C. R. Mec.*, vol. 338, no.6, pp. 311–320, 2010.

[25] N. Ghadermarzy, Y. Plan, and O. Yilmaz, "Near-optimal sample complexity for convex tensor completion". *Inf. Inference*, vol. 8, no. 3, pp. 577–619, 2019.

[26] S. Friedland and L. H. Lim. "Nuclear norm of higher-order tensors," *Math. Comput.*, vol. 87, no. 311, pp. 1255–1281, 2018.

[27] L. R. Tucher, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[28] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do, "Efficient tensor completion for color image and video recovery: low-rank tensor train," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2466–2479, 2017.

[29] J. Liu, P. Musialski, P. Wonka, and J. Ye. "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, 2013.

[30] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Alg. Appl.*, vol. 435, no. 3, pp. 641–658, 2011.

[31] 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 Anal. Appl.*, vol. 34, no. 1, pp. 148–172, 2013.

[32] Z. M. Zhang, G. Ely, S. Aeron, N. Hao, and M. E. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2014, pp. 3842–3849.

[33] 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 Process.*, vol. 23, no. 4, pp. 1678–1693, 2014.

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

[35] Z. M. Zhang and S. Aeron, "Exact tensor completion using t-SVD," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1511–1526, 2017.

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

[37] J. F. Cai, J. Y. Li, and D. Xia, "Generalized low-rank plus sparse tensor estimation by fast Riemannian optimization," *J. Am. Stat. Assoc.*, 2022: 1–17.

[38] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," *Ann. Statist.*, vol. 50, no. 1, pp. 1–29, 2022.

[39] A. Aidini, G. Tsagkatakis, and P. Tsakalides, "Quantized tensor robust principal component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 2453–2457.

[40] A. Aidini, G. Tsagkatakis, and P. Tsakalides, "1-bit tensor completion," *Electron Imag.*, vol. 13, pp. 1–6, 2018.

[41] N. Ghadermarzy, Y. Plan, and O. Yilmaz, "Learning tensors from partial binary measurements," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 29–40, 2018.

[42] B. H. Li, X. N. Zhang, X. L. Li, and H. C. Lu, "Tensor completion from one-bit observations," *IEEE Trans. Image Process.*, vol. 28, no.1, pp. 170–180, 2019.

[43] J. Y. Hou, F. Zhang, Y. Wang, and J. J. Wang, "Robust Low-tubal-rank Tensor Recovery from Binary Measurements," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4355–4373, 2022.

[44] J. Y. Hou, F. Zhang, J. J. Wang, "One-bit tensor completion via transformed tensor singular value decomposition," *Appl. Math. Model.*, vol. 95, pp. 760–782, 2021.

[45] C. W. Lee and M. Y. Wang. "Tensor denoising and completion based on ordinal observations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5778–5788.

[46] P. Zhou, C. Y. Lu, Z. C. Lin, and C. Zhang, "Tensor factorization for low-rank tensor completion," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1152–1163, 2017.

[47] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1665–1697, 2012.

[48] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[49] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[50] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[51] M. Ledoux and M. Talagrand. *Probability in Banach space: Isoperimetry and Processes*. Springer Science & Business Media, Berlin, 2013.

[52] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[53] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1–2, pp. 459–494, 2014.
[54] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 272–279.
[55] Y. Xu and W. Yin, "A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion," Tech. Rep., Rice University CAAM, Sep. 2012.
[56] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.
[57] J. Y. Hou, J. J. Wang, F. Zhang, and J. W. Huang, "One-bit compressed sensing via $\ell_p(0 < p < 1)$-minimization method," *Inverse Probl.*, vol. 36, no. 5, 2020, Art. no. 055005.
[58] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic Range and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, 2010.
[59] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[60] 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. Conf. Comput. Vis.*, 2001, pp. 416–423.
[61] D. Martin, R. Shafer, and B. LaRue, "An order-p tensor factorization with applications in imaging," *SIAM J. Sci. Comput.*, vol. 35, no. 1, pp. A474–A490, 2013.
[62] W. J. Qin, H. L. Wang, F. Zhang, J. J. Wang, X. Luo, and T. W. Huang, "Low-rank high-order tensor completion with applications in visual data," *IEEE Trans. Image Process.*, vol. 31, pp. 2433–2448, 2022.

Jianjun Wang (Member, IEEE) received the B.Sc. degree in mathematical education and the M.Sc. degree in fundamental mathematics from Ningxia University, Yinchuan, China, in 2000 and 2003, respectively, and the Ph.D. degree in applied mathematics from the Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, in 2006. He is currently a Professor with the School of Mathematics and Statistics, Southwest University, Chongqing, China. His research interests include machine learning, data mining, neural networks, and sparse learning.

Jingyao Hou received the B.Sc. degree in mathematics and applied mathematics from the School of Mathematics and Statistics, Chongqing Three Gorges University, Chongqing, China, in 2014, and the M.Sc. and Ph.D. degrees in statistics from the School of Mathematics and Statistics, Southwest University, Chongqing, China, in 2017 and 2021, respectively. He is currently a Research Associate at the College of Mathematics & Information, China West Normal University, Nanchong, China. His research focuses on one-bit compressed sensing and tensor sparsity.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degrees in physics and electrical engineering both from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002.

She was a Professor with the Department of Electrical Engineering at Technion and was a Visiting Professor at Stanford. She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She is also a Visiting Professor at MIT, a Visiting Scientist at the Broad Institute, and an Adjunct Professor at Duke University. She is the author of the book Sampling Theory: Beyond Bandlimited Systems and the coauthor of five other books published by the Cambridge University Press. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging, and optics.