# Sparse and Low-rank Tensor Estimation via Cubic Sketchings

**Botao Hao**
Princeton University

**Anru Zhang**
University of Wisconsin-Madison

**Guang Cheng**
Purdue University

## Abstract

In this paper, we propose a general framework for sparse and low-rank tensor estimation from cubic sketchings. A two-stage non-convex implementation is developed based on sparse tensor decomposition and thresholded gradient descent, which ensures exact recovery in the noiseless case and stable recovery in the noisy case with high probability. The non-asymptotic analysis sheds light on an interplay between optimization error and statistical error. The proposed procedure is shown to be rate-optimal under certain conditions. As a technical by-product, novel high-order concentration inequalities are derived for studying high-moment sub-Gaussian tensors. An interesting tensor formulation illustrates the potential application to high-order interaction pursuit in high-dimensional linear regression.

## 1 Introduction

The rapid advance in modern scientific technology gives rise to a wide range of high-dimensional tensor data (Kroonenberg, 2008; Kolda and Bader, 2009). Accurate estimation and fast communication/processing of tensor-valued parameters are crucially important in practice. For example, a tensor-valued predictor, which characterizes the association between brain diseases and scientific measurements, such as magnetic resonance imaging, becomes the point of interest (Zhou et al., 2013; Li et al., 2018; Sun and Li, 2017). Another example is tensor-valued image acquisition algorithms that can considerably reduce the number of required samples by exploiting the compressibility property of signals (Caiafa and Cichocki, 2013; Friedland et al., 2014).

In particular, the following tensor estimation model is widely considered in recent literatures,

$$y_i = \langle \mathscr{T}^*, \mathscr{X}_i \rangle + \epsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Here, $\mathscr{X}_i$ and $\epsilon_i$ are the measurement tensor and noise, respectively. The goal is to estimate the unknown tensor $\mathscr{T}^*$ from measurements $\{y_i, \mathscr{X}_i\}_{i=1}^n$. A number of specific settings with varying forms of $\mathscr{X}_i$ have been studied, e.g., tensor completion (Liu et al., 2013; Yuan and Zhang, 2016, 2017; Zhang, 2019; Montanari and Sun, 2018), tensor regression (Zhou et al., 2013; Li et al., 2018; Raskutti et al., 2018; Chen et al., 2016; Li and Zhang, 2017; Sun and Li, 2017), multi-task learning (Romera-Paredes et al., 2013), etc.

In this paper, we focus on the case that the measurement tensor can be written in a cubic form, i.e., $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$. The cubic sketching form of $\mathscr{X}_i$ is motivated by interaction effect estimation. High-dimensional high-order interaction models have been considered under a variety settings (Bien et al., 2013; Hao and Zhang, 2014; Fan et al., 2016; Basu et al., 2018). By writing $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$, we find that the interaction model has an interesting tensor representation which allows us to estimate high-order interaction terms using tensor techniques. This is in contrast with the existing literature that mostly focused on pair-wise interactions due to the model complexity and computational difficulties. More detailed discussions will be provided in Section A.

In practice, the total number of measurements $n$ is considerably smaller than the number of parameters

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

in unknown tensor $\mathscr{T}^*$, due to all kinds of restrictions such as time and storage. Fortunately, a variety of high-dimensional tensor data possess intrinsic structures, such as low-rankness (Kolda and Bader, 2009) and sparsity (Sun et al., 2017), which highly reduce the effective dimension of the parameter and make the accurate estimation possible. Please refer to Section 3 for low-rank and sparse assumptions.

In this paper, we propose a computationally efficient non-convex optimization approach for sparse and low-rank tensor estimation via cubic-sketchings. Our procedure is two-stage: (i) obtain an initial estimate via the method of tensor moment (motivated by high-order Stein's identity), and then apply sparse tensor decomposition to the initial estimate to output a provably warm start; (ii) use a thresholded gradient descent to iteratively refine the warm start along each tensor mode until convergence.

In theory, we carefully characterize the optimization and statistical errors at each iteration step. The output estimate is shown to converge in a geometric rate to an estimation with minimax optimal rate in statistical error (in terms of tensor Frobenius norm). In particular, after a logarithmic factor of iterations, whenever $n \gtrsim K^2(s\log(ep/s))^{\frac{3}{2}}$, the proposed estimator $\widehat{\mathscr{T}}$ achieves

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \leq C\sigma^2 \frac{Ks\log(p/s)}{n} \qquad (1.2)$$

with high probability, where $n$, $s$, $K$, $p$, and $\sigma^2$ are the number of measurements, the sparsity (an absolute number of non-zeros), rank, dimension, and noise level, respectively. We further establish the matching minimax lower bound to show that (1.2) is indeed optimal over a large class of sparse low-rank tensors. Our optimality result can be further extended to the non-sparse domain (such as tensor regression (Chen et al., 2016; Rauhut et al., 2017)) – to the best of our knowledge, this is the first optimality result in both sparse and non-sparse low-rank tensor regressions.

The above theoretical analyses are non-trivial due to the non-convexity of the empirical risk function, and the need to develop some new high-order sub-Gaussian concentration inequalities. Specifically, the empirical risk function in consideration satisfies neither restricted strong convexity (RSC) condition nor sparse eigenvalue (SE) condition in general. Thus, many previous results, such as the one based on local optima analysis (Wang et al., 2014; Loh and Wainwright, 2015; Chen et al., 2016), are not directly applicable. Moreover, the structure of cubic-sketching tensor leads to high-order products of sub-Gaussian random variables. Thus, the matrix analysis based on Hoeffding-type or Bernstein-type concentration inequality (Cai and Zhang, 2015; Chen et al., 2015) will lead to sub-optimal statistical rate and sample complexity. This motivates us to develop new high-order concentration inequalities and sparse tensor-spectral-type bound, i.e., Lemmas 1 and 8. These new technical results are obtained based on the careful partial truncation of high-order products of sub-Gaussian random variables and the argument of bounded $\psi_\alpha$-norm (Adamczak et al., 2011), and may be of independent interest.

A related line of research is low-rank matrix estimation in the literature, e.g., the spectral method and nuclear norm minimization (Candès and Recht, 2009; Keshavan et al., 2010; Koltchinskii et al., 2011). However, our cubic sketching model is by-no-means a simple extension from matrix estimation problems. In general, many related concepts or methods for matrix data, such as singular value decomposition, are problematic to apply in the tensor framework (Richard and Montanari, 2014; Zhang and Xia, 2018). It is also found that simple unfolding or matricizing of tensors may lead to suboptimal results due to the loss of structural information (Mu et al., 2014). Technically, the tensor nuclear norm is NP-hard to even approximate (Yuan and Zhang, 2016, 2017; Friedland and Lim, 2018), and thus the method to handle tensor low-rankness is particularly different from the matrix.

## 2 Preliminary

For any set $A$, let $|A|$ be the cardinality. The $\mathrm{diag}(\boldsymbol{x})$ is a diagonal matrix generated by $\boldsymbol{x}$. For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{x} \circ \boldsymbol{y}$ is the outer product. Define $\|\boldsymbol{x}\|_q := (|x_1|^q + \cdots + |x_p|^q)^{1/q}$. Let $\boldsymbol{e}_j$ be the canonical vectors, whose $j$-th entry equals to 1 and all other entries equal to zero. We next introduce notations and operations on the matrix. For matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J] \in \mathbb{R}^{I\times J}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L] \in \mathbb{R}^{K\times L}$, the *Kronecker product* is defined as a $(IK)$-by-$(JL)$ matrix $\boldsymbol{A}\otimes\boldsymbol{B} = [\boldsymbol{a}_1\otimes\boldsymbol{B} \ \cdots \boldsymbol{a}_J\otimes\boldsymbol{B}]$, where $\boldsymbol{a}_j\otimes\boldsymbol{B} = (a_{j1}\boldsymbol{B}^\top, \ldots, a_{jI}\boldsymbol{B}^\top)^\top$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns $J = L$, the *Khatri-Rao product* is defined as $\boldsymbol{A} \odot \boldsymbol{B} = [\boldsymbol{a}_1 \circ \boldsymbol{b}_1, \boldsymbol{a}_2 \circ \boldsymbol{b}_2, \cdots, \boldsymbol{a}_J \circ \boldsymbol{b}_J] \in \mathbb{R}^{IK \times J}$. If the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the same dimension, the *Hadamard product* is their element-wise matrix product, such that $(\boldsymbol{A} * \boldsymbol{B})_{ij} = \boldsymbol{A}_{ij} \cdot \boldsymbol{B}_{ij}$.

In the end, we focus on tensor notation and relevant operations. Suppose $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is an order-3 tensor, then the $(i,j,k)$-th element of $\mathcal{X}$ is denoted by $[\mathcal{X}]_{ijk}$. The successive tensor multiplication with vectors $\boldsymbol{u} \in \mathbb{R}^{p_2}$, $\boldsymbol{v} \in \mathbb{R}^{p_3}$ is denoted by $\mathcal{X} \times_2 \boldsymbol{u} \times_3 \boldsymbol{v} = \sum_{j \in [p_2], l \in [p_3]} u_j v_l \mathcal{X}_{[:,j,l]} \in \mathbb{R}^{p_1}$. We say $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is *rank-one* if it can be written as the outer product of three vectors, i.e., $\mathcal{X} = \boldsymbol{x}_1 \circ \boldsymbol{x}_2 \circ \boldsymbol{x}_3$ or $[\mathcal{X}]_{ijk} = x_{1i} x_{2j} x_{3k}$ for all $i, j, k$.

More generally, we may decompose a tensor as the sum of rank one tensors as follows,

$$\mathcal{X} = \sum_{k=1}^{K} \eta_k \boldsymbol{x}_{1k} \circ \boldsymbol{x}_{2k} \circ \boldsymbol{x}_{3k}, \tag{2.1}$$

where $\eta_k \in \mathbb{R}, \boldsymbol{x}_{1k} \in \mathbb{S}^{p_1 - 1}, \boldsymbol{x}_{2k} \in \mathbb{S}^{p_2 - 1}, \boldsymbol{x}_{3k} \in \mathbb{S}^{p_3 - 1}$. This is the so-called CANDECOMP/PARAFAC, or CP decomposition (Kolda and Bader, 2009) with CP-rank being defined as the minimum number $K$ such that (2.1) holds. Several tensor norms also need to be introduced. The tensor Frobenius norm and tensor spectral norm are defined respectively as

$$\|\mathcal{X}\|_F = \sqrt{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} \mathcal{X}_{ijk}^2}, \tag{2.2}$$

$$\|\mathcal{X}\|_{op} := \sup_{\boldsymbol{u} \in \mathbb{R}^{p_1}, \boldsymbol{v} \in \mathbb{R}^{p_2}, \boldsymbol{w} \in \mathbb{R}^{p_3}} \frac{|\langle \mathcal{X}, \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w} \rangle|}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2 \|\boldsymbol{w}\|_2},$$

where $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i,j,k} \mathcal{X}_{ijk} \mathcal{Y}_{ijk}$. Clearly, $\|\mathcal{X}\|_F^2 = \langle \mathcal{X}, \mathcal{X} \rangle$. We also consider the following sparse tensor spectral norm,

$$\|\mathcal{X}\|_s := \sup_{\substack{\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = \|\boldsymbol{c}\| = 1 \\ \max\{\|\boldsymbol{a}\|_0, \|\boldsymbol{b}\|_0, \|\boldsymbol{c}\|_0\} \le s}} \left| \langle \mathcal{X}, \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c} \rangle \right|. \tag{2.3}$$

By definition, $\|\mathcal{X}\|_s \le \|\mathcal{X}\|_{op}$.

## 3 Symmetric Tensor Estimation via Cubic Sketchings

In this section, we focus on the estimation of sparse and low-rank symmetric tensors,

$$\begin{aligned} y_i &= \langle \mathscr{T}^*, \mathscr{X}_i \rangle + \epsilon_i, \\ \mathscr{X}_i &= \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i \in \mathbb{R}^{p \times p \times p}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.1}$$

where $\boldsymbol{x}_i$ are random vectors with i.i.d. standard normal entries. As previously discussed, the tensor parameter $\mathscr{T}^*$ often satisfies certain low-dimensional structures in practice, among which the factor-wise sparsity and low-rankness (Raskutti et al., 2018) commonly appear. We thus assume $\mathscr{T}^*$ is CP rank-$K$ for $K \ll p$ and the corresponding factors are sparse, $\mathscr{T}^* = \sum_{k=1}^K \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$, where $\|\boldsymbol{\beta}_k^*\|_2 = 1, \|\boldsymbol{\beta}_k^*\|_0 \le s, \forall k \in [K]$. The CP low-rankness has been widely assumed in literature for its nice scalability and simple formulation (Li and Li, 2010; Li and Zhang, 2017; Sun and Li, 2017).

Based on observations $\{y_i, \mathscr{X}_i\}_{i=1}^n$, we propose to estimate $\mathscr{T}^*$ via minimizing the empirical squared loss since the closed form gradient provides computational convenience: $\widehat{\mathscr{T}} = \text{argmin}_{\mathscr{T}} \mathcal{L}(\mathscr{T})$, subject to $\mathscr{T}$ is sparse and low-rank, where

$$\begin{aligned} \mathcal{L}(\mathscr{T}) &= \mathcal{L}(\eta_k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \frac{1}{n} \sum_{i=1}^n (y_i - \langle \mathscr{T}, \mathscr{X}_i \rangle)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left( y_i - \sum_{k=1}^K \eta_k \left( \boldsymbol{x}_i^\top \boldsymbol{\beta}_k \right)^3 \right)^2. \end{aligned} \tag{3.2}$$

Equivalently, (3.2) can be written as,

$$\begin{aligned} &\min_{\eta_k, \boldsymbol{\beta}_k} \frac{1}{n} \sum_{i=1}^n \Big( y_i - \sum_{k=1}^K \eta_k (\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)^3 \Big)^2, \\ &\text{s.t. } \|\boldsymbol{\beta}_k\|_2 = 1, \|\boldsymbol{\beta}_k\|_0 \le s, \ \text{for } k \in [K]. \end{aligned} \tag{3.3}$$

Clearly, (3.3) is a non-convex optimization problem. To solve it, we propose a two-stage method as described in the next two subsections.

### 3.1 Initialization

Due to the non-convex optimization (3.3), a straightforward implementation of many local search algorithms, such as gradient descent and alternating minimization, may easily get trapped into local optimums and obtain sub-optimal statistical performances. Inspired by recent advances of spectral method (e.g., EM algorithm (Zhang et al., 2016), phase retrieval (Cai et al., 2016), and tensor SVD (Zhang and Xia, 2018)), we propose to evaluate an initial estimate $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ via the method of moment and sparse tensor decomposition (a variant of high-order spectral method) in the following Steps 1 and 2, respectively. The pseudo-code is given in Algorithm 1.

**Step 1: Unbiased Empirical Moment Estimator.** Construct the empirical moment based estimator $\mathcal{T}_s := \frac{1}{6}[\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i - \sum_{j=1}^p (\boldsymbol{m}_1 \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{m}_1 \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{m}_1)]$, where $\boldsymbol{m}_1 := \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i$, $\boldsymbol{e}_j$ is the canonical vector.

As will be shown in Lemma 3, $\mathcal{T}_s$ is an unbiased estimator of $\mathscr{T}^*$. The construction is motivated by high-order Stein's identity (Janzamin et al. (2014); also see Theorem 5 for a complete statement). Intuitively speaking, based on the third-order score function for a Gaussian random vector $\boldsymbol{x}$: $\mathcal{S}_3(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{x} \circ \boldsymbol{x} - \sum_{j=1}^p (\boldsymbol{x} \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{x} \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{x})$, we can construct the unbiased estimator of $\mathscr{T}^*$ by properly choosing a continuously differentiable function in high-order Stein's identity. See the proof of Lemma 3 for more details.

**Step 2: Sparse Tensor Decomposition.** The method of moment estimator obtained in Step 1 provides an initial estimate for tensor $\mathscr{T}^*$. Then we further obtain good initialization for the factors $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ via truncation and alternating rank-1 power iterations (Anandkumar et al., 2014; Sun et al., 2017), $\mathcal{T}_s \approx \sum_{k=1}^K \eta_k^{(0)} \boldsymbol{\beta}_k^{(0)} \circ \boldsymbol{\beta}_k^{(0)} \circ \boldsymbol{\beta}_k^{(0)}$. Note that the tensor power iterations recover one rank-1 component per time. To identify all rank-1 components, we generate a large number of different initialization vectors at first, implement a clustering step, and choose the centroids as the estimates in the initialization stage.

More specifically, we firstly choose a large integer $M \gg K$ and generate $M$ starting vectors $\{\boldsymbol{b}_m^{(0)}\}_{m=1}^M \in \mathbb{R}^p$ through sparse SVD as described in Algorithm 3 (described in the supplementary). Then for each $\boldsymbol{b}_m^{(0)}$, we apply the following truncated power update:

$$\begin{aligned} \widetilde{\boldsymbol{b}}_m^{(l+1)} &= \frac{\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}}{\|\mathcal{T}_s \times_2 \boldsymbol{b}_m^{(l)} \times_3 \boldsymbol{b}_m^{(l)}\|_2}, \\ \boldsymbol{b}_m^{(l+1)} &= \frac{T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})}{\|T_d(\widetilde{\boldsymbol{b}}_m^{(l+1)})\|_2}, \quad l = 0, \dots, \end{aligned} \tag{3.4}$$

where $\times_2, \times_3$ are tensor multiplication operators defined in Section 2 and $T_d(\boldsymbol{x}) \in \mathbb{R}^p$ is a truncation operator that sets all but the largest $d$ entries in absolute values to zero for any vector $\boldsymbol{x} \in \mathbb{R}^p$. We run power iterations till its convergence, and denote $\boldsymbol{b}_m$ as the outcome. Finally, we apply $K$-means to partition $\{\boldsymbol{b}_m\}_{m=1}^M$ into $K$ clusters, then let the centroids of the output clusters be $\{\boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$ and calculate $\eta_k^{(0)} = \mathcal{T}_s \times_1 \boldsymbol{\beta}_k^{(0)} \times_2 \boldsymbol{\beta}_k^{(0)} \times_3 \boldsymbol{\beta}_k^{(0)}$ for $k \in [K]$.

**Algorithm 1** Initialization in cubic sketchings

```
Require: response $\{y_i\}_{i=1}^n$, sketching vector $\{\boldsymbol{x}_i\}_{i=1}^n$,
    truncation level $d$, rank $K$, stopping error $\epsilon = 10^{-4}$.
Step 1: Calculate the moment-based tensor $\mathcal{T}_s$.
Step 2:
  For $m = 1$ to $M$
    Generate $\boldsymbol{b}_m^{(0)}$ through Algorithm 3.
    Repeat power update (3.4).
    Until $\|\boldsymbol{b}_m^{(l+1)} - \boldsymbol{b}_m^{(l)}\|_2 \le \epsilon$.
  End for.
  Perform $K$-means for $\{\boldsymbol{b}_m^{(l)}\}_{m=1}^M$. Denote the centroids of $K$ clusters by $\{\boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.
  Calculate $\eta_k^{(0)} = \mathcal{T}_s \times_1 \boldsymbol{\beta}_k^{(0)} \times_2 \boldsymbol{\beta}_k^{(0)} \times_3 \boldsymbol{\beta}_k^{(0)}, k \in [K]$.
return symmetric tensor estimator $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.
```

### 3.2 Thresholded Gradient Descent

After obtaining a warm start in the first stage, we propose to apply the thresholding gradient descent to iteratively refine the solution to the non-convex optimization problem (3.3). Specifically, denote $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathbb{R}^{p \times n}$, $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)^\top \in \mathbb{R}^K$ and $\boldsymbol{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p \times K}$. Recall that $\mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = \mathcal{L}(\mathscr{T})$, and hence let

$$\nabla_{\boldsymbol{B}} \mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = (\nabla_{\boldsymbol{\beta}_1} \mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})^\top, \dots, \nabla_{\boldsymbol{\beta}_K} \mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})^\top),$$

be the gradient function with respect to $\boldsymbol{B}$. Based on the detailed calculation in Lemma S.1, $\nabla_{\boldsymbol{B}} \mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta})$ can be written as

$$\begin{aligned} \nabla_{\boldsymbol{B}} \mathcal{L}(\boldsymbol{B}, \boldsymbol{\eta}) = &\frac{6}{n}[\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^3 \boldsymbol{\eta} - \boldsymbol{y}]^\top \\ &\cdot [(\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^2 \odot \boldsymbol{\eta}^\top)^\top \odot \boldsymbol{X}]^\top, \end{aligned} \tag{3.5}$$

where $\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^3$ and $\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^2$ are entry-wise cubic and squared matrices of $(\boldsymbol{B}^\top \boldsymbol{X})^\top$. Define $\varphi_h(x)$ as the thresholding function with a level $h$ that satisfies the following minimal assumptions: $|\varphi_h(x) - x| \le h, \forall x \in \mathbb{R}$, and $\varphi_h(x) = 0$, when $|x| \le h$. Many widely used thresholding schemes, such as hard thresholding $H_h(x) = xI_{(|x|>h)}$, soft-thresholding $S_h(x) = \text{sign}(x)\max(|x| - h, x)$, satisfy the above assumption. With slightly abuse of notations, we further define the vector thresholding function as $\varphi_h(\boldsymbol{x}) = (\varphi_h(x_1), \dots, \varphi_h(x_p))$, for $\boldsymbol{x} \in \mathbb{R}^p$.

The initial estimates $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{B}^{(0)}$ will be updated by thresholded gradient descent in two steps summarized in Algorithm 2. It is noteworthy that only $\boldsymbol{B}$ is updated in the Step 3, while $\boldsymbol{\eta}$ will be updated in Step 4 after the update of $\boldsymbol{B}$ is finished.

**Step 3: Updating $\boldsymbol{B}$ via Thresholded Gradient descent.** We update $\boldsymbol{B}^{(t)}$ in each iteration step via thresholded gradient descent,

$$\mathrm{vec}(\boldsymbol{B}^{(t+1)}) = \varphi_{\frac{\mu \boldsymbol{h}(\boldsymbol{B}^{(t)})}{\phi}}(\mathrm{vec}(\boldsymbol{B}^{(t)}) - \frac{\mu}{\phi}\nabla_{\boldsymbol{B}}\mathcal{L}(\boldsymbol{B}^{(t)}, \boldsymbol{\eta}^{(0)})).$$

Here, $\mu$ is the step size and $\phi = \sum_{i=1}^{n} y_i^2/n$ serves as an approximation for $(\sum_{k=1}^{K}\eta_k^*)^2$ (see Lemma 15); $\boldsymbol{h}(\boldsymbol{B}) \in \mathbb{R}^{1\times K}$ is the thresholding level defined as

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{B}) =& \sqrt{\frac{4\log np}{n^2}}[\{\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^3 \boldsymbol{\eta}^{(0)} - \boldsymbol{y}\}^2]^\top \\ & \cdot \{\{(\boldsymbol{B}^\top \boldsymbol{X})^\top\}^2 \odot \boldsymbol{\eta}^{(0)\top}\}^2.\end{aligned}$$

**Step 4: Updating $\boldsymbol{\eta}$ via Normalization.** We normalize each column of $\boldsymbol{B}^{(T)}$ and estimate the weight parameter as

$$\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)^\top = \Big(\frac{\boldsymbol{\beta}_1^{(T)}}{\|\boldsymbol{\beta}_1^{(T)}\|_2}, \ldots, \frac{\boldsymbol{\beta}_K^{(T)}}{\|\boldsymbol{\beta}_K^{(T)}\|_2}\Big),$$

$$\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_K)^\top = \Big(\eta_1^{(0)}\|\boldsymbol{\beta}_1^{(T)}\|_2^3, \ldots, \eta_K^{(0)}\|\boldsymbol{\beta}_K^{(T)}\|_2^3\Big)^\top.$$

The final estimator for $\mathscr{T}^*$ is $\widehat{\mathscr{T}} = \sum_{k=1}^{K}\widehat{\eta}_k\widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k$.

**Algorithm 2** Thresholded gradient descent in cubic sketchings

```
Require: response $\{y_i\}_{i=1}^n$, sketching vector $\{\boldsymbol{x}_i\}_{i=1}^n$, step size $\mu$, rank $K$, stopping error $\epsilon = 10^{-4}$, warm-start $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.
1: Step 3: Let $t = 0$.
2:    Repeat Compute thresholding level $\boldsymbol{h}(\boldsymbol{B})$ and calculate the thresholded gradient descent update.
3:    Until $\|\boldsymbol{B}^{(T+1)} - \boldsymbol{B}^{(T)}\|_F \le \epsilon$.
4: Step 4: Perform column-wise normalization and update the weight. Construct the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \widehat{\eta}_k\widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k \circ \widehat{\boldsymbol{\beta}}_k$.
5: return symmetric tensor estimator $\widehat{\mathscr{T}}$.
```

**Algorithm 3** Sparse SVD

```
Require: tensor $\mathcal{T}_s$, cardinality parameter $d$.
1: Compute $\widetilde{\boldsymbol{\theta}} = T_d(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim \mathcal{N}(0, I_d)$.
2: Calculate $\boldsymbol{u}$ as the leading singular vector of $\mathcal{T}_s \times_1 \widetilde{\boldsymbol{\theta}}$.
3: return the sparse vector $T_d(\boldsymbol{u})/\|\boldsymbol{u}\|_2$.
```

# 4 Theoretical Analysis

In this section, we establish the geometric convergence rate in optimization error and minimax optimal rate in statistical error of the proposed symmetric tensor estimator.

## 4.1 Assumptions

Conditions 1-3 are on the true tensor parameter $\mathscr{T}^*$ while Conditions 4-5 are on the measurement scheme. The first condition guarantees the model identifiability for CP-decomposition.

**Condition 1** (Uniqueness of CP-decomposition). The CP-decomposition form is unique in the sense that if there exists another CP-decomposition $\mathscr{T}^* = \sum_{k=1}^{K'}\eta_k^{*'}\boldsymbol{\beta}_k^{*'} \circ \boldsymbol{\beta}_k^{*'} \circ \boldsymbol{\beta}_k^{*'}$, it must have $K = K'$ and be invariant up to a permutation of $\{1, \ldots, K\}$.

For technical purpose, we introduce the following conditions to ensure that the CP-decomposition of $\mathscr{T}^*$ has a regular form in the sense that the operator norm of $\mathscr{T}^*$ can be bounded by the largest factor and all factors are in the same order. Similar assumptions were previously used in literature (e.g., Zhou et al. (2013); Sun et al. (2017)).

**Condition 2** (Parameter space). The CP-decomposition $\mathscr{T}^* = \sum_{k=1}^{K}\eta_k^*\boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$ satisfies

$$\begin{aligned}&\|\mathscr{T}^*\|_{op} \le C\eta_{\max}^*, \quad K = \mathcal{O}(s), \\ &\text{and} \quad R = \eta_{\max}^*/\eta_{\min}^* \le C',\end{aligned} \tag{4.1}$$

for some absolute constants $C, C'$, where $\eta_{\min}^* = \min_k \eta_k^*$ and $\eta_{\max}^* = \max_k \eta_k^*$. Recall that $s$ is the sparsity for $\boldsymbol{\beta}_k^*$.

The performance of Step 2, i.e. the tensor decomposition for initialization, is crucial to the final estimation. However, as shown in the seminal work of Håstad (1990); Hillar and Lim (2013), the estimation of the low-rank tensor is NP-hard in general. Hence, we impose the following incoherence condition that is widely used in tensor decomposition literature (Anandkumar et al., 2014; Sun et al., 2017).

**Condition 3** (Parameter incoherence). The true tensor components are incoherent such that

$$\Gamma := \max_{1\le k_1\ne k_2\le K}|\langle\boldsymbol{\beta}_{k_1}^*, \boldsymbol{\beta}_{k_2}^*\rangle| \le \min\{C'' K^{-\frac{3}{4}}R^{-1}, s^{-\frac{1}{2}}\},$$

where $R$ is the singular value ratio defined in (4.1) and $C''$ is some small constant.

We also introduce the following conditions on noise and sample complexity.

**Condition 4** (Sub-exponential noise). The noise $\{\epsilon_i\}_{i=1}^n$ are i.i.d. randomly generated with mean 0 and variance $\sigma^2$ satisfying $0 < \sigma < C\sum_{k=1}^K \eta_k^*$. $(\epsilon_i/\sigma)$ is sub-exponential distributed, i.e., there exists constant $C_\epsilon > 0$ such that $\|(\epsilon_i/\sigma)\|_{\psi_1} := \sup_{p\geq 1} p^{-1}(\mathbb{E}|\epsilon_i/\sigma|^p)^{1/p} \leq C_\epsilon$, and independent of $\{\mathscr{X}_i\}_{i=1}^n$.

The sample complexity condition is crucial for our algorithm, especially in the initialization stage. Ignoring any polylog factors, Condition 5 is even weaker than the sparse matrix estimation case ($n \gtrsim s^2$) in Cai et al. (2016).

**Condition 5** (Sample complexity). We assume a sufficient number of observations is observed, $n \geq C^{'''} K^2 (s\log(ep/s))^{\frac{3}{2}} \log^4 n$.

### 4.2 Main Theoretical Results

Our main Theorem 1 shows that based on a good initializer, the output from the proposed thresholded gradient descent can achieve optimal statistical rate after sufficient iterations. Here, we define a contraction parameter $0 < \kappa = 1 - 32\mu K^{-2} R^{-\frac{8}{3}} < 1$, and also denote $\mathcal{E}_1 = 4K\eta_{\max}^{*\frac{2}{3}}\varepsilon_0^2$ and $\mathcal{E}_2 = C_0 \eta_{\min}^{*-\frac{4}{3}}/16$ for some $C_0 > 0$.

**Theorem 1** (Statistical Error and Optimization Error). Suppose Conditions 3-5 hold and the initial estimator $\{\boldsymbol{\beta}_k^{(0)}, \eta_k^{(0)}\}_{k=1}^K$ satisfies

$$\max_{1\leq k\leq K} \Big\{ \|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*| \Big\} \lesssim K^{-1}, \quad (4.2)$$

with probability at least $1 - \mathcal{O}(1/n)$ and $|\text{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$. Assume the step size $\mu \leq \mu_0$, where $\mu_0$ is defined in (S.6). Then, the output from the thresholded gradient descent update satisfies:

- For any $t = 0, 1, 2, \ldots$, the factor-wise estimator satisfies

  $$\begin{aligned}&\sum_{k=1}^K \Big\| \sqrt[3]{\eta_k^{(0)}}\boldsymbol{\beta}_k^{(t+1)} - \sqrt[3]{\eta_k^*}\boldsymbol{\beta}_k^* \Big\|_2^2 \\ \leq &\mathcal{E}_1 \kappa^t + \mathcal{E}_2 \frac{\sigma^2 s \log p}{n},\end{aligned} \quad (4.3)$$

  with probability at least $1 - \mathcal{O}(tKs/n)$.

- When the total number of iterations is no smaller than

  $$T^* = \Big( \log(\frac{n}{\sigma^2 s \log p} \vee 1) + \log \frac{\mathcal{E}_1}{\mathcal{E}_2} \Big) / \log \kappa^{-1}, (4.4)$$

  there exists a constant $C_1$ (independent of $K, s, p, n, \sigma^2$) s.t. the final estimator $\widehat{\mathscr{T}} = \sum_{k=1}^K \eta_k^{(0)} \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)} \circ \boldsymbol{\beta}_k^{(T^*)}$ is upper bounded by

  $$\Big\| \widehat{\mathscr{T}} - \mathscr{T}^* \Big\|_F^2 \leq \frac{C_1 \sigma^2 K s \log p}{n}, \quad (4.5)$$

  with probability at least $1 - \mathcal{O}(T^* Ks/n)$.

**Remark 1** . From (4.3), the error bound can be decomposed into an optimization error $\mathcal{E}_1\kappa^t$ (which decays with a geometric rate as iterations) and a statistical error $\mathcal{E}_2 \frac{\sigma^2 s \log p}{n}$ (which does not decay as iterations). In particular, the convergence rate of the optimization error relies on the rank $K$ and the singular value ratio $R$ in the sense that the smaller $K$ or $R$, the faster convergence. Also from (4.5), we note that in the special case that $\sigma = 0$, $\widehat{\mathscr{T}}$ exactly recover $\mathscr{T}^*$ with high probability.

The next theorem shows that Steps 1 and 2 of Algorithm 1 provides a good initializer required in Theorem 1.

**Theorem 2** (Initialization Error). Suppose the number of initializations $L \geq K^{C_3\gamma^{-4}}$, where $\gamma$ is a constant defined in (S.3). Given that Conditions 1-4 hold, the initial estimator obtained from Steps 1-2 with a truncation level $s \leq d \leq Cs$ satisfies

$$\begin{aligned}&\max_{1\leq k\leq K} \Big\{ \|\boldsymbol{\beta}_k^{(0)} - \boldsymbol{\beta}_k^*\|_2, |\eta_k^{(0)} - \eta_k^*| \Big\} \\ \leq &C_2 K R \delta_{n,p,s} + \sqrt{K}\Gamma^2,\end{aligned} \quad (4.6)$$

and $|\text{supp}(\boldsymbol{\beta}_k^{(0)})| \lesssim s$ with probability at least $1 - 5/n$, where

$$\delta_{n,p,s} = (\log n)^3 \Big( \sqrt{\frac{s^3 \log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}} \Big).$$

Moreover, if the sample complexity condition 5 is satisfied, then the above bound satisfies (4.2).

**Remark 2** (Interpretation of initialization error). The upper bound of (4.6) consists of two terms, which corresponds to the approximation error of $\mathcal{T}_s$ to $\mathscr{T}^*$ and the incoherence condition of $\boldsymbol{\beta}_k^*$'s, respectively. Especially, the former converges to zero as

$n$ grows while the latter does not. This indicates that the convergence rate of the initial estimate is significantly slower than that of the final estimate after iterative updates, unless $n \gtrsim (s\log(ep/s))^2$ and $\Gamma^2 \lesssim \sqrt{\frac{s\log(ep/s)}{nK}}$. More detailed numerical comparisons will be provided later in Section 5.

The proof of Theorems 1 and 2 are involved and postponed to Section S.I-S.II in the supplementary materials. The combination of Theorems 1 and 2 immediately yields the following upper bound for the final estimate as one main result in this paper.

**Theorem 3** (Upper Bound). Suppose Conditions 1 – 5 hold, $s \le d \le Cs$. After $T^*$ iterations, there exists a constant $C_1$ not depending on $K, s, p, n, \sigma^2$, such that the proposed procedure yields

$$\left\|\widehat{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \le \frac{C_1\sigma^2 K s\log p}{n}, \tag{4.7}$$

with probability at least $1 - \mathcal{O}(T^*Ks/n)$, where $T^*$ is defined in (4.4).

The above upper bound turns out to match with the minimax lower bound for a large class of sparse and low rank tensors.

**Theorem 4** (Lower Bound). Consider the following class of sparse and low-rank tensors,

$$\mathcal{F}_{p,K,s} = \left\{ \mathscr{T} : \begin{array}{l} \mathscr{T} = \sum_{k=1}^K \eta_k \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \\ \|\boldsymbol{\beta}_k\|_0 \le s, \text{ for } k \in [K], \\ \mathscr{T} \text{ satisfies Conditions 1, 2, and 3.} \end{array} \right\}. \tag{4.8}$$

Suppose that $\{\mathscr{X}_i\}_{i=1}^n$ are i.i.d standard normal cubic sketchings with i.i.d. $N(0,\sigma^2)$ noise in (3.1). We have the following lower bound result,

$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T}\in\mathcal{F}_{p,K,s}} \mathbb{E}\left\|\widetilde{\mathscr{T}} - \mathscr{T}\right\|_F^2 \ge c\sigma^2\frac{Ks\log(ep/s)}{n}.$$

The proof of Theorem 4 is deferred to Section S.III in the supplementary materials. Combining Theorem 3 and Theorem 4 together, we immediately obtain the following minimax-optimal rate for sparse and low-rank tensor estimation with cubic sketchings when $\log p \asymp \log(p/s)$:

$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T}^*\in\mathcal{F}_{p,K,s}} \mathbb{E}\left\|\widetilde{\mathscr{T}} - \mathscr{T}^*\right\|_F^2 \asymp \sigma^2\frac{Ks\log(p/s)}{n}. \tag{4.9}$$

The rate in (4.9) sheds light upon the effect of dimension $p$, noise level $\sigma^2$, sparsity $s$, sample size $n$ and rank $K$ to the estimation performance.

**Remark 3** . We would like to highlight our algorithmic and theoretical results automatically hold for non-sparse case with all the truncation/thresholding steps removed. If no sparsity assumption and $n \ge p^{3/2}$, one can apply similar arguments of Theorems 1-3 to show that the output estimate satisfies optimal rate $\sqrt{\sigma^2 Kp/n}$ in terms of tensor Frobenius norm. Our analysis does not take advantage of sparsity assumption.

### 4.3 Key Lemmas: High-order Concentration Inequalities

As mentioned earlier, one major challenge for theoretical analysis of cubic sketching is to handle heavy tails of high-order Gaussian moments. One can only handle up-to second moments of sub-Gaussian random variables by directly applying the existing Hoeffding's or Bernstein's concentration inequalities. Rather, we need to develop the following high-order concentration inequality as technical tools. It provides a generic spectral-type concentration inequality that can be used to quantify the approximation error for $\mathcal{T}_s$ introduced in Step 1 of the proposed procedure. The proof of 1 is given in Section S.II.

**Lemma 1** (Concentration inequality for Gaussian cubic sketchings). Suppose $\{\boldsymbol{x}_{1i}\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_{p_1})$, $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$ are fixed vectors. Define $M = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i}, \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \rangle \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i} \circ \boldsymbol{x}_{1i}$. Then $\mathbb{E}(M) = 6\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_1 + 3\sum_{m=1}^p (\boldsymbol{\beta}_1 \circ \boldsymbol{e}_m \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{\beta}_1 \circ \boldsymbol{e}_m + \boldsymbol{e}_m \circ \boldsymbol{e}_m \circ \boldsymbol{\beta}_1)$, and

$$\begin{aligned} &\left\|M - \mathbb{E}(M)\right\|_s \\ &\le C(\log n)^3\left(\sqrt{\frac{s^3\log^3(ep/s)}{n^2}} + \sqrt{\frac{s\log(ep/s)}{n}}\right)\|\boldsymbol{\beta}_1\|_2^3, \end{aligned}$$

with probability at least $1 - 10/n^3$. Here, $C$ is an absolute constant and $\|\cdot\|_s$ is the sparse tensor spectral norm defined in (2.3).

## 5 Numerical Results

In this section, we empirically examine the effect of noise level, CP-rank, sample size, dimension, and sparsity on the estimation performance. In each setting, we generated $\mathscr{T}^* = \sum_{k=1}^K \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*$, where $|\text{supp}(\boldsymbol{\beta}_k^*)| = s$ was uniformly selected from $\{1, \ldots, p\}$, the nonzero entries of $\boldsymbol{\beta}_k^*$ were drawn

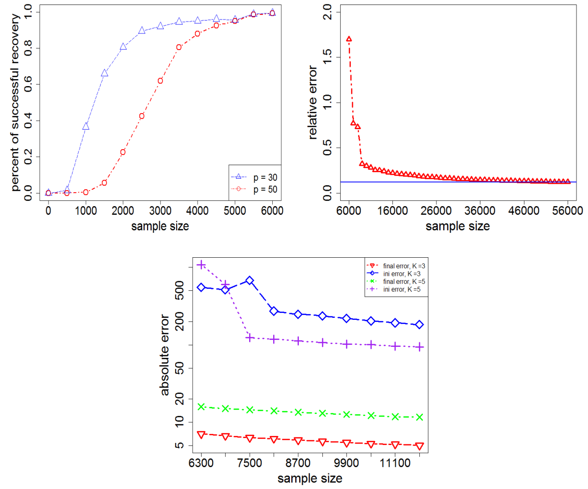

Figure 1: Percent of successful recovery with varying sample size (top left panel). Log absolute estimation error of initial estimation error (top right panel) and initialization/final estimation error comparisons (bottom panel).

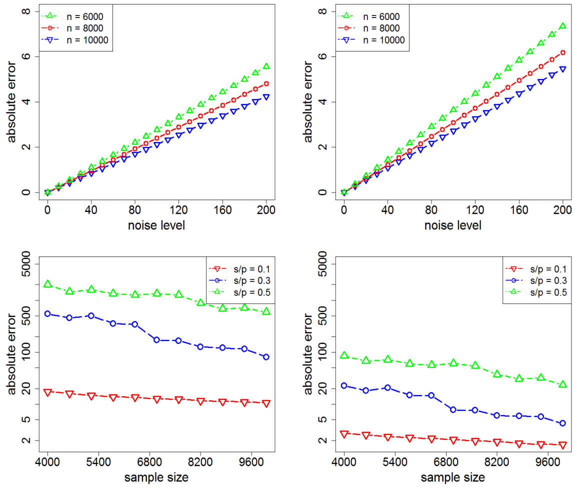

Figure 2: Estimation error for different sample sizes and noise levels. The top left panel is $p = 30$ and the top right panel is $p = 50$. The bottom left panel is for initial estimation error and the bottom right panel is for final estimation error.

from standard Gaussian distribution. Next we normalized each vector $\boldsymbol{\beta}_k^*$ and aggregated the coefficient as $\eta_k^*$. The cubic sketchings $\{\mathscr{X}_i\}_{i=1}^n$ were generated as $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$, where $\{\boldsymbol{x}_i\}_{i=1}^n$ were from standard Gaussian distribution. The noise $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma^2)$.

First, we consider the percent of successful recovery in the noiseless case. Let $K = 3$, $s/p = 0.3$, $p = 30$ or $50$, so that the total number of unknown parameters in $\mathscr{T}^*$ is $2.7 \times 10^4$ or $1.25 \times 10^5$. The sample size $n$ ranges from 500 to 6000. The recovery is called successful if the relative error $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F / \|\mathscr{T}^*\|_F < 10^{-4}$. We report the percent of successful recovery in Figure 1. It is clear from Figure 1 that the empirical relation with dimensionality and sample size is consistent with our theory.

We then move to the noisy case where the empirical estimation error is examined. We select $K = 3$, $s/p = 0.3$, $p = 30$ or $50$, $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma^2)$ and consider two specific scenarios: (1) sample size $n$ = 6000, 8000, or 10000, $s/p = 0.3$, the noise level $\sigma$ varies from 0 to 200; (2) noise level $\sigma = 200$, sample size $n$ varies from 4000 to 10000, $p = 30$, $s/p = 0.1, 0.3, 0.5$. The estimation errors in terms of $\|\widehat{\mathscr{T}} - \mathscr{T}^*\|_F$ under these two scenarios are plotted in Figures 2, respectively. From these results, we can see that the proposed algorithm achieves reasonable estimation performance: Algorithms 1 and 2 yield more accurate estimation with smaller variance $\sigma^2$ and/or large value of sample size $n$.

Next, we compare the estimation errors of initial and final estimators for different ranks and sample sizes. First we set $K = 3, p = 30, s/p = 0.3$ and consider the noiseless setting. It is clear from Figure 1 that the initialization error decays sufficiently, but does not converge to zero as sample size $n$ grows. This result matches our theoretical findings in Theorem 2. After sufficient steps of thresholded gradient descent (Steps 3 and 4 in Algorithm 2), the initial estimator is refined to lead to the final estimate that is proven to be minimax-optimal. Thus, we evaluate and compare estimation errors for both initial and final estimators for $K = 3$ or 5 and growing sample sizes $n$. We can see from the bottom panel of Figure 1, the final estimator is more stable and accurate compared with the initial one, which illustrates the merit of thresholded gradient descent step of the proposed procedure.

**Acknowledgements**

Guang Cheng would like to acknowledge support by NSF DMS-1712907, DMS-1811812, DMS-1821183, and Office of Naval Research (ONR N00014- 18-2759). While completing this work, Guang Cheng was a member of Institute for Advanced Study, Princeton and visiting Fellow of SAMSI for the Deep Learning Program in the Fall of 2019; he would like to thank both Institutes for their hospitality. Anru Zhang would like to acknowledge support by NSF CAREER-1944904, NSF DMS-1811868, and NIH R01 GM131399.

## References

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88.

Anandkumar, A., Ge, R., and Janzamin, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bogucki, R. (2015). Suprema of canonical weibull processes. *Statistics & Probability Letters*, 107:253–263.

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

Cai, T. T. and Zhang, A. (2015). Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138.

Caiafa, C. F. and Cichocki, A. (2013). Multidimensional compressed sensing and their applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(6):355–380.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Chen, H., Raskutti, G., and Yuan, M. (2016). Nonconvex projected gradient descent for generalized low-rank tensor regression. *arXiv preprint arXiv:1611.10349*.

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

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*.

Friedland, S., Li, Q., and Schonfeld, D. (2014). Compressive sensing of sparse tensors. *IEEE Transactions on Image Processing*, 23(10):4438–4447.

Friedland, S. and Lim, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Håstad, J. (1990). Tensor rank is np-complete. *Journal of algorithms (Print)*, 11(4):644–654.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45.

Hitczenko, P., Montgomery-Smith, S., and Oleszkiewicz, K. (1997). Moment inequalities for sums of certain independent symmetric random variables. *Studia Math*, 123(1):15–42.

Hung, H., Lin, Y.-T., Chen, P., Wang, C.-C., Huang, S.-Y., and Tzeng, J.-Y. (2016). Detection of genegene interactions using multistage sparse and low-rank regression. *Biometrics*, 72(1):85–94.

Janzamin, M., Sedghi, H., and Anandkumar, A. (2014). Score function features for discriminative learning: matrix and tensor framework. *arXiv preprint arXiv:1412.2863*.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998.

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51:455–500.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. Wiley Series in Probability and Statistics.

Ledoux, M. (2005). *The concentration of measure phenomenon*. Number 89. American Mathematical Soc.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, pages 1–16.

Li, N. and Li, B. (2010). Tensor completion for onboard compression of hyperspectral images. In *2010 IEEE International Conference on Image Processing*, pages 517–520. IEEE.

Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545.

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:208–220.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Montanari, A. and Sun, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 73–81, Bejing, China. PMLR.

Nguyen, N. H., Drineas, P., and Tran, T. D. (2015). Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229.

Raskutti, G., Yuan, M., and Chen, H. (2018). Convex regularization for high-dimensional multi-response tensor regression. *The Annals of Statistics*, to appear.

Rauhut, H., Schneider, R., and Stojanac, Ž. (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262.

Richard, E. and Montanari, A. (2014). A statistical model for tensor pca. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2897–2905. Curran Associates, Inc.

Romera-Paredes, B., Aung, M. H., Bianchi-Berthouze, N., and Pontil, M. (2013). Multilinear multitask learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1444–III–1452. JMLR.org.

Sidiropoulos, N. D. and Kyrillidis, A. (2012). Multiway compressed sensing for sparse low-rank tensors. *IEEE Signal Processing Letters*, 19(11):757–760.

Stein, C., Diaconis, P., Holmes, S., Reinert, G., et al. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pages 1–25. Institute of Mathematical Statistics.

Sun, W. W. and Li, L. (2017). Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944.

Sun, W. W., Lu, J., Liu, H., and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916.

Talagrand, M. (1994). The supremum of some canonical processes. *American Journal of Mathematics*, 116(2):283–325.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2015). Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*.

Vershynin, R. (2012). *Compressed sensing*, chapter Introduction to the non-asymptotic analysis of random matrices, pages 210–268. Cambridge Univ. Press.

Wang, Z., Liu, H., and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of statistics*, 42(6):2164.

Yu, B. (1997). Assouad, fano, and le cam. *Festschrift for Lucien Le Cam*, 423:435.

Yuan, M. and Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068.

Yuan, M. and Zhang, C.-H. (2017). Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766.

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

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2016). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552.