The Annals of Statistics
2021, Vol. 49, No. 1, 76–99
https://doi.org/10.1214/20-AOS1942

STATISTICALLY OPTIMAL AND COMPUTATIONALLY EFFICIENT LOW RANK TENSOR COMPLETION FROM NOISY ENTRIES

BY DONG XIA[1], MING YUAN[2] AND CUN-HUI ZHANG[3]

[1]*Department of Mathematics, Hong Kong University of Science and Technology, ming.yuan@columbia.edu*
[2]*Department of Statistics, Columbia University, ming.yuan@columbia.edu*
[3]*Department of Statistics and Biostatistics, Rutgers University, czhang@stat.rutgers.edu*

In this article, we develop methods for estimating a low rank tensor from noisy observations on a subset of its entries to achieve both statistical and computational efficiencies. There have been a lot of recent interests in this problem of noisy tensor completion. Much of the attention has been focused on the fundamental computational challenges often associated with problems involving higher order tensors, yet very little is known about their statistical performance. To fill in this void, in this article, we characterize the fundamental statistical limits of noisy tensor completion by establishing minimax optimal rates of convergence for estimating a kth order low rank tensor under the general ℓ_p $(1 \le p \le 2)$ norm which suggest significant room for improvement over the existing approaches. Furthermore, we propose a polynomial-time computable estimating procedure based upon power iteration and a second-order spectral initialization that achieves the optimal rates of convergence. Our method is fairly easy to implement and numerical experiments are presented to further demonstrate the practical merits of our estimator.

1. Introduction.

Let $\mathbf{T} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ be a kth order tensor, or multilinear array. In the noisy tensor completion problem, we are interested in recovering $\mathbf{T}$ from observations of a subset of its entries. More specifically, our sample consists of n independent copies $\{(Y_i, \omega_i) : 1 \le i \le n\}$ of a random pair (Y, ω) obeying

$$Y = T(\omega) + \xi, \tag{1.1}$$

where ω is uniformly sampled from $[d_1] \times \cdots \times [d_k]$ where $[d] = \{1, 2, \dots, d\}$, and independent of the measurement error ξ that is assumed to be a centered sub-Gaussian random variable. Of particular interest here is the high dimensional settings where the sample size n may be much smaller than the ambient dimension $d_1 \cdots d_k$. In this case, it may not be possible to estimate an arbitrary kth order tensor well but it is possible to do so if we focus on tensors that resides in a manifold of lower dimension in $\mathbb{R}^{d_1 \times \cdots \times d_k}$. A fairly general and practically appropriate example is the class of tensors of low rank. Problems of this type arise naturally in a wide range of applications including imaging and computer vision (e.g., [19, 21, 31]), signal processing (e.g., [17, 20, 23, 25]), latent variable modeling (e.g., [1, 6, 8, 30]), to name a few. Although many statistical methods and algorithms have been proposed for these problems, very little is known about their theoretical properties and to what extent they work and may not work.

An exception is the special case of matrices, that is, $k = 2$, for which low rank completion from noisy entries is well understood; see, for example, [5, 13, 14, 16, 24] and references

Received November 2017; revised August 2019.
MSC2020 subject classifications. 60K35.
Key words and phrases. Tensor completion, low-rank, power iteration, spectral initialization, minimax optimality.

therein. In particular, as shown by [16], an estimator based on nuclear norm regularization, denoted by $\widehat{\mathbf{T}}^{\mathrm{KLT}}$, converges to $\mathbf{T}$ at the rate of

$$\frac{\|\widehat{\mathbf{T}}^{\mathrm{KLT}}-\mathbf{T}\|_{\ell_2}}{(d_1d_2)^{1/2}}=O_p\biggl((\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)\sqrt{\frac{r(d_1\vee d_2)\log(d_1\vee d_2)}{n}}\biggr),\tag{1.2}$$

where $a\vee b=\max\{a,b\}$, and $\|\cdot\|_{\ell_p}$ $(p\geq 1)$ denotes the vectorized ℓ_p norm. Note that the dimension of the manifold of rank r matrices is of the order $r(d_1+d_2)$, the aforementioned convergence rate is therefore expected to be optimal, up to the logarithmic factor. Indeed a rigorous argument was given by [16] to show that it is optimal, up to the logarithmic factor, in the minimax sense. In contrast to the matrix case, our understanding of higher order tensors ($k\geq 3$) is fairly limited. The main goal of the current work is to fill in this void by establishing minimax lower bounds for estimating $\mathbf{T}$, and developing computationally efficient methods that attain the optimal statistical performance.

Treatment of higher order tensors poses several fundamental challenges. On the one hand, many of the basic tools and properties for matrices, particularly those pertaining to low rank approximation, are no longer valid for higher order tensors. For example, many of the aforementioned estimating procedures developed for matrices are based on singular value decompositions (SVD) whose generalization to tensors, however, is rather delicate. A particularly popular generalization of SVD to tensors is the so-called higher order singular value decomposition (HOSVD). The (truncated) HOSVD to tensors, generally, do not deliver a good (far from being the best) low rank approximation. Additionally, unlike SVD, HOSVD is bitterly statistically unreliable due to the noise accumulation from the larger dimensions. As shown in Proposition 1, the naive HOSVD requires a significantly large sample size for noisy tensor completion. As a result, although many of these approaches have been extended to higher order tensors in recent years, their theoretical properties remain largely unclear. And recent studies on a related problem, namely nuclear norm minimization for exact tensor completion without noise, point to many fundamental differences between matrices and higher order tensors despite their superficial similarities; see, for example, [32, 33]. On the other hand, as pointed out by [11], most computational problems related to higher order tensors, including the simple task of evaluating tensor spectral and nuclear norms, are typically NP-hard. Consequently, convex relaxation approaches by nuclear norm regularizations, while being attractive for matrices, are computationally intractable for tensors. This dictates that it is essential to take computational efficiency into account in devising statistically optimal estimating procedures for $\mathbf{T}$. We note that the classical one-step MLE can easily achieve statistically optimal convergence rates, which is, however, computationally infeasible in general. The real difficulty in noisy tensor completion is to gain statistical optimality and computational efficiency simultaneously.

Because of these difficulties, results for higher order noisy tensor completion comparable to (1.2) are scarce. The strongest result to date is due to [4]. They focused on the case of third order tensors, that is $k=3$, and proved that, under suitable conditions which we shall discuss in details later on, there is a polynomial-time computable estimator, denoted by $\widehat{\mathbf{T}}^{\mathrm{BM}}$, such that

$$\frac{\|\widehat{\mathbf{T}}^{\mathrm{BM}}-\mathbf{T}\|_{\ell_1}}{d_1d_2d_3}=O_p\biggl(\frac{\|\mathbf{T}\|_*(d_{\min}d_{\max}^2)^{1/4}\log^2(d_{\max})}{\sqrt{n}}+\frac{\|\Xi\|_{\ell_1}}{d_1d_2d_3}+\frac{\|\Xi\|_{\ell_\infty}}{\sqrt{n}}\biggr),\tag{1.3}$$

where $d_{\min}=d_1\wedge d_2\wedge d_3$, $d_{\max}=d_1\vee d_2\vee d_3$, $\|\cdot\|_*$ stands for the tensor nuclear norm, and Ξ is a $d_1\times d_2\times d_3$ random tensor whose entries are independent copies of ξ. More recently, the authors of [22] considered approximating a general kth order $d\times\cdots\times d$ cubic tensor in the absence of noise, that is, $\Xi = 0$, and proposed a spectral method that yield an estimator $\widehat{\mathbf{T}}^{\mathrm{MS}}$ obeying

$$\|\widehat{\mathbf{T}}^{\mathrm{MS}} - \mathbf{T}\|_{\ell_2} = O_p\biggl(\frac{\|\mathbf{T}\|_{\ell_2} r^{1/3} d^{k/6} \log^3(d)}{n^{1/3}}\biggr), \tag{1.4}$$

under more restrictive assumptions, where r is the rank of $\mathbf{T}$. Clearly, the bounds (1.3) and (1.4) are statistically suboptimal in view of the degrees of freedom. However, it remains unknown to what extent these bounds can be improved, especially if we take computational efficiency into account. The present article addresses this question specifically, and provides a definitive answer.

In particular, we investigate the minimax optimal estimates for a low rank tensor under the general ℓ_p $(1 \le p \le 2)$ loss. We propose a computational efficient procedure based on low rank projection of an unbiased estimate of $\mathbf{T}$, and show that, if $\mathbf{T}$ is well conditioned, then the estimation error of the resulting estimate, denoted by $\widehat{\mathbf{T}}$, satisfies

$$\biggl(\frac{1}{d_1 \cdots d_k}\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_p}^p\biggr)^{1/p} = O_p\biggl((\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)\sqrt{\frac{r d_{\max}\log(d_{\max})}{n}}\biggr), \tag{1.5}$$

provided that

$$n \ge C\bigl(r^{(k-1)/2}(d_1 \cdots d_k)^{1/2}\log^{k+2}(d_{\max}) + r^{k-1} d_{\max}\log^{2(k+2)}(d_{\max})\bigr), \tag{1.6}$$

for a suitable constant $C > 0$. Here, $d_{\max} = d_1 \vee \cdots \vee d_k$ as before. The above result continues to hold if r represents the maximum multilinear rank of $\mathbf{T}$. Note that under the typical incoherent assumptions for $\mathbf{T}$, $\|\mathbf{T}\|_* \ge \|\mathbf{T}\|_{\ell_2} \asymp (d_1 \cdots d_k)^{1/2}\|\mathbf{T}\|_{\ell_\infty}$ as we shall discuss in further details later. Therefore, $\widehat{\mathbf{T}}$ converges to $\mathbf{T}$ at a much faster rate than both $\widehat{\mathbf{T}}^{\mathrm{BM}}$ and $\widehat{\mathbf{T}}^{\mathrm{MS}}$. Furthermore, we show that the rates given by (1.5) are indeed minimax optimal, up to the logarithmic factor, over all incoherent tensors of low multilinear rank.

Our estimator also has its practical appeal when compared with earlier proposals. In general, computing the best low rank approximation to a large tensor is rather difficult; see, for example, [10] and [11]. The root cause of the problem is the highly nonconvex nature of the underlying optimization problem. As a result, there could be exponentially many local optima (see, e.g., [2, 3]). To address this challenge, we devise a strategy that first narrows down the search area for the best low rank approximation using a novel spectral method and then applies power iteration to identify a local optimum within the search area. The high level idea of combining spectral method and power iterations to yield improved estimate is similar in spirit to the classical one-step MLE. Existing polynomial-time computable estimators such as $\widehat{\mathbf{T}}^{\mathrm{BM}}$ often involve solving huge semidefinite programs which are known not to scale well for large problems. In contrast, our approach is not only polynomial-time computable but also very easy to implement.

It is worth pointing out that, in order to achieve the minimax optimal rate of convergence given by (1.5), a sample size requirement (1.6) is imposed. This differs from the matrix case and appears to be inherent to tensor related problems. More specifically, it was argued in [4] that, if there is no polynomial-time algorithm for refuting random 3-SAT of d variables with $d^{3/2-\epsilon}$ clauses for any $\epsilon > 0$, then any polynomial-time computable estimator of a $d \times d \times d$ tensor $\mathbf{T}$ is inconsistent whenever $n = O(d^{3/2-\epsilon})$. This suggests that the sample size requirement of the form (1.6) is likely necessary for any polynomial-time computable estimator because thus far, indeed there is no polynomial-time algorithm for refuting random 3-SAT of d variables with $o(d^{3/2})$ clauses in spite of decades of pursuit. These observations from [4] point to a fundamental difference between matrix and tensor completions with respect to the statistical and computational gaps in sample size requirements. In matrix completion, the sample size $n = O(d)$ is sufficient and necessary for both the statistical consistency and computational efficiency. There is no gap in sample size requirements between the statistical and computational aspects. Intriguingly, such a gap exists in higher order tensor completion. Indeed, a sample size $n = O(d)$ suffices to guarantee the statistically consistent recovery (e.g., MLE) of a $d \times d \times d$ tensor. In contrary, a sample size $n = O(d^{3/2})$ is necessary to devise computationally tractable estimators to reconstruct the tensor consistently.

Our work here is also related to a fast growing literature on exact low rank tensor completion where we observe the entries without noise, that is $\xi = 0$ in (1.1), and aim to recover $\mathbf{T}$ perfectly; see, for example, [12, 28, 32, 33] and references therein. The two types of problems, albeit connected, have many fundamental differences which manifest prominently in their respective treatment. On the one hand, the noisy completion considered here is technically more involved because of the presence of measurement error ξ. In fact, much of our analysis is devoted to carefully control the adverse effect of ξ. On the other hand, our interest in the noisy case is in seeking a good estimate or approximation of $\mathbf{T}$, which is to be contrast with the noiseless case where the goal is for exact recovery and, therefore, more difficult to achieve. As we shall demonstrate later, this distinction allows for simpler algorithms and sharper analysis in the noisy setting.

The rest of the paper is organized as follows. We first describe the proposed estimation method in Section 2. Some useful spectral bounds are given in Section 3 which we shall use to establish theoretical properties for our estimator in Section 4. Numerical experiments are presented in Section 5 to complement our theoretical results. Proofs of the main results are presented in Section 6.

2. Methodology.

We are interested in estimating a tensor $\mathbf{T} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ based on observations $\{(Y_i, \omega_i) : 1 \le i \le n\}$ that follow

$$Y_i = T(\omega_i) + \xi_i, \quad i = 1, \dots, n,$$

assuming that $\mathbf{T}$ is of low rank. To this end, we first review some basic notions and facts about tensors.

2.1. *Background and notation.*

Recall that the mode-j fibers of a kth order tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ are the d_j dimensional vectors

$$\{A(i_1, \dots, i_{j-1}, \cdot, i_{j+1}, \dots, i_k) : i_1 \in [d_1], \dots, i_k \in [d_k]\},$$

that is, vectors obtained by fixing all but the jth indices of $\mathbf{A}$. Let $\mathcal{M}_j(\mathbf{A})$ be the mode-j flattening of $\mathbf{A}$ so that $\mathcal{M}_j(\mathbf{A})$ is a $d_j \times (d_1 \cdots d_{j-1} d_{j+1} \cdots d_k)$ matrix whose column vectors are the mode-j fiber of $\mathbf{A}$. For example, for third order tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, define the matrix $\mathcal{M}_1(\mathbf{A}) \in \mathbb{R}^{d_1 \times (d_2 d_3)}$ by the entries

$$\mathcal{M}_1(\mathbf{A})\bigl(i_1, (i_2 - 1)d_3 + i_3\bigr) = A(i_1, i_2, i_3), \quad \forall i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3].$$

Denote by $r_j(\mathbf{A})$ the rank of $\mathcal{M}_j(\mathbf{A})$. The tuple $(r_1(\mathbf{A}), \dots, r_k(\mathbf{A}))$ is often referred to the multilinear ranks of $\mathbf{A}$. Denote by

$$r_{\max}(\mathbf{A}) := \max\{r_1(\mathbf{A}), \dots, r_k(\mathbf{A})\}.$$

Another common notion of tensor rank is the so-called canonical polyadic (CP) rank. Recall that a rank-one tensor can be expressed as

$$\mathbf{A} = u_1 \otimes \cdots \otimes u_k,$$

for some $u_j \in \mathbb{R}^{d_j}$. Here, the $\otimes$ stands for the outer product so that

$$(u_1 \otimes \cdots \otimes u_k)(i_1, \dots, i_k) = u_1(i_1) \cdot \cdots \cdot u_k(i_k).$$

The CP rank, or sometimes rank for short, of a tensor $\mathbf{A}$, denoted by $R(\mathbf{A})$, is defined as the smallest number of rank-one tensors needed to sum up to $\mathbf{A}$. It is common in the literature to refer to a tensor as low rank if its CP rank is small. It is clear that $r_{\max}(\mathbf{A}) \le R(\mathbf{A}) \le r_1(\mathbf{A}) \cdots r_k(\mathbf{A})/r_{\max}(\mathbf{A})$ so that a tensor of low CP rank necessarily has small multilinear ranks. We shall focus on tensors with low multilinear ranks in the rest of the paper because of this connection between the two notions of tensor ranks.

Suppose that we know a priori that $\mathbf{T}$ is of low rank. A natural starting point for estimating $\mathbf{T}$ is the least squares estimate:

$$\begin{aligned}&\min_{\mathbf{A}\in\Theta(r_1,\ldots,r_k)}\left\{\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - A(\omega_i)\bigr)^2\right\}\\ &\quad = \min_{\mathbf{A}\in\Theta(r_1,\ldots,r_k)}\left\{\frac{1}{n}\sum_{i=1}^{n}A^2(\omega_i) - \frac{2}{n}\sum_{i=1}^{n}Y_iA(\omega_i)\right\},\end{aligned}$$

where

$$\Theta(r_1,\ldots,r_k) := \{\mathbf{A}\in\mathbb{R}^{d_1\times\cdots\times d_k} : r_k(\mathbf{A}) \le r_k\}$$

is the collection of kth order tensors whose multilinear ranks are upper bounded by $r_1,\ldots,r_k$, respectively. Note that similarly defined least squares estimator for tensors with small CP rank may not be well defined (see, e.g., [10]). Noting that ω_is are uniformly sampled from $[d_1]\times\cdots[d_k]$, we shall replace the first term on the right-hand side simply by its population counterpart $\|\mathbf{A}\|_{\ell_2}^2$, leading to

$$\begin{aligned}&\min_{\mathbf{A}\in\Theta(r_1,\ldots,r_k)}\left\{(d_1\cdots d_k)^{-1}\|\mathbf{A}\|_{\ell_2}^2 - \frac{2}{n}\sum_{i=1}^{n}Y_iA(\omega_i)\right\}\\ &\quad = \min_{\mathbf{A}\in\Theta(r_1,\ldots,r_k)}\left\{\left\|\frac{d_1\cdots d_k}{n}\sum_{i=1}^{n}Y_i\mathbf{e}_{\omega_i} - \mathbf{A}\right\|_{\ell_2}^2\right\}.\end{aligned}$$

Here, $\mathbf{e}_\omega$ is a tensor whose entries are zero except that its ωth entry is one. In other words, we can estimate $\mathbf{T}$ by the best multilinear ranks-$(r_1,\ldots,r_k)$ approximation of

$$\widehat{\mathbf{T}}^{\text{init}} := \frac{d_1\cdots d_k}{n}\sum_{i=1}^{n}Y_i\mathbf{e}_{\omega_i}. \tag{2.1}$$

A similar approach can also be applied to more general sampling schemes, and was first introduced by [16] in the matrix setting. However, there is a major challenge when doing so for higher order tensors: computing low rank approximations to a higher order ($k \ge 3$) tensor is NP-hard in general (see, e.g., [11]). This makes it practically less meaningful to study the properties of the exact projection of $\widehat{\mathbf{T}}^{\text{init}}$ onto $\Theta(r_1,\ldots,r_k)$. To overcome this hurdle, we adapted the power iteration as a way to compute an "approximate" projection. We shall show in later sections that, even though it may not produce the true projection, running the algorithm for a sufficiently large but finite number of iterations is guaranteed to yield an estimate that attains the minimax optimal rate of convergence.

2.2. *Estimation via power iterations.* Recall that we are interested in solving

$$\min_{\mathbf{A}\in\Theta(r_1,\ldots,r_k)}\{\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{A}\|_{\ell_2}^2\} \tag{2.2}$$

The objective function is smooth so that many smooth optimization algorithms might be employed. In particular, we shall consider using power iterations, one of the most common methods for low rank approximation; see, for example, [15].

Algorithm 1 Power Iterations

```
Input: $\widehat{\mathbf{T}}^{\text{init}}$, $U_j^{(0)}$ for $j = 1, 2, \dots, k$, and parameter $\text{iter}_{\max}$.
Output: $\widehat{\mathbf{T}}$.
Set counter $\text{iter} = 0$.
while $\text{iter} < \text{iter}_{\max}$ do
    Set $\text{iter} = \text{iter} + 1$ and $j = 0$.
    while $j < k$ do
        Set $j = j + 1$.
        Set $U_j^{(\text{iter})}$ to be the first $r_j$ left singular vectors of
            $\mathcal{M}_j(\widehat{\mathbf{T}}^{\text{init}} \times_{j'<j} U_{j'}^{(\text{iter})} \times_{j'>j} U_{j'}^{(\text{iter}-1)})$.
    end while
end while
Return $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}^{\text{init}} \times_{j=1}^{k} U_j^{(\text{iter})}(U_j^{(\text{iter})})^\top$.
```

For a tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, denote by U_j the left singular vectors of $\mathcal{M}_j(\mathbf{A})$. Then we can find a tensor $\mathbf{C} \in \mathbb{R}^{r_1(\mathbf{A}) \times \cdots \times r_k(\mathbf{A})}$ such that

$$\mathbf{A} = \mathbf{C} \times_1 U_1^\top \times_2 \cdots \times_k U_k^\top.$$

Here, the marginal multiplication $\times_j$ between a tensor $\mathbf{A}$ and a matrix B of conformable dimension results in a tensor whose entries are defined as

$$(\mathbf{A} \times_j B)(i_1, \dots, i_{j-1}, i_j, i_{j+1}, \dots, i_k) = \sum_{i'_j} A(i_1, \dots, i_{j-1}, i'_j, i_{j+1}, \dots, i_k) B(i'_j, i_j).$$

In particular, it can be derived that, if $\mathbf{A}$ is the solution to (2.2), then

$$\mathbf{C} = \widehat{\mathbf{T}}^{\text{init}} \times_{j=1}^{k} U_j,$$

and U_j is a $d_j \times r_j$ matrix whose columns are the leading singular vectors of

$$\mathcal{M}_j(\widehat{\mathbf{T}}^{\text{init}} \times_{j' \neq j} U_{j'}).$$

This naturally leads to Algorithm 1 which updates $\mathbf{C}$ and U_js in an iterative fashion.

The power iteration as described above is guaranteed to converge for any given initial value $U_j^{(0)}$s. But it is only guaranteed to converge to a local optimum of (2.2); see, for example, [26] and references therein for further discussion about the convergence of power method.

2.3. *Spectral initialization.* Because of the highly nonconvex nature of the space of low rank tensors, the local convergence of Algorithm 1 may not ensure that it yields a good estimate. For example, as shown by [3], there could be exponentially many (in ds) local optima and vast majority of these local optima are far from the best low rank approximation. See also [2]. An observation key to our development is that if we start from an initial value not too far from the global optimum, then a local optimum reached by these locally convergent algorithms may be as good an estimator as the global optimum. In fact, as we shall show later, if we start from an appropriate initial value, then even running Algorithm 1 for a finite number of iteration could yield a high quality estimate of $\mathbf{T}$.

It turns out, however, that the construction of an initial value for U_js that are both close to the truth, that is, the leading left singular vectors of $\mathcal{M}_j(\mathbf{T})$, and polynomial-time computable is a fairly challenging task. An obvious choice is to start with higher order singular

Algorithm 2 Compute Estimate of $\mathbf{T}$ from $\{(Y_i, \omega_i) : 1 \le i \le n\}$

```
Input: Observations $\{(Y_i, \omega_i) : 1 \le i \le n\}$, threshold $\lambda$, and parameter $\mathrm{iter}_{\max}$.
Output: $\widehat{\mathbf{T}}$.
Compute $\widehat{\mathbf{T}}^{\mathrm{init}}$ as given by (2.1).
Initialize $U_j$s:
for $j = 1, \ldots k$ do
    Compute $\widehat{\mathbf{N}}_j$ as given by (2.3).
    Compute the eigenvectors, denoted by $U_j^{(0)}$, of $\widehat{\mathbf{N}}_j$ with eigenvalue greater than $\lambda^2$.
end for
Run Algorithm 1 with inputs $\widehat{\mathbf{T}}^{\mathrm{init}}$, $U_j^{(0)}$s and $\mathrm{iter}_{\max}$ to get $\widehat{\mathbf{T}}$.
Return $\widehat{\mathbf{T}}$.
```

value decomposition (HOSVD; see, e.g., [9]), and initialize U_j as the left singular vectors of $\mathcal{M}_j(\widehat{\mathbf{T}}^{\mathrm{init}})$. It is clear that how close such an initialization is to the truth is determined by the difference $\mathcal{M}_j(\widehat{\mathbf{T}}^{\mathrm{init}}) - \mathcal{M}_j(\mathbf{T})$. This approach, however, neglects the fact that we are only interested in the left singular vectors of a potentially very "fat" $(d_j \ll (d_1 \cdots d_k)/d_j)$ matrix. As a result, it can be shown that an unnecessarily large amount of samples are needed to ensure that such an initialization is sufficiently "close" to the truth.

To overcome this difficulty, we adopt a second-order spectral method developed by [28]. Note that the column vectors of left singular vectors of $\mathcal{M}_j(\mathbf{T})$ are also the leading eigenvectors of

$$\mathbf{N}_j := \mathcal{M}_j(\mathbf{T})\mathcal{M}_j(\mathbf{T})^\top.$$

Therefore, we could consider estimating the eigenvectors of $\mathbf{N}_j$ instead. Specifically, we first estimate $\mathbf{N}_j$ by the following U-statistic:

$$\widehat{\mathbf{N}}_j := \frac{(d_1 \cdots d_k)^2}{n(n-1)} \sum_{1 \le i \ne i' \le n} Y_i Y_{i'} \mathcal{M}_j(\mathbf{e}_{\omega_i}) \mathcal{M}_j(\mathbf{e}_{\omega_{i'}})^\top. \tag{2.3}$$

We then initialize U_j as the collection of eigenvectors of $\widehat{\mathbf{N}}_j$ with eigenvalues greater than a threshold λ to be specified later. We shall show later that $\widehat{\mathbf{N}}_j$ concentrates around $\mathbf{N}_j$ better than $\mathcal{M}_j(\widehat{\mathbf{T}}^{\mathrm{init}})$ around $\mathcal{M}_j(\mathbf{T})$ and, therefore, allows for better initialization of U_js. In summary, our estimating procedure can be described by Algorithm 2.

We now turn our attention to the theoretical properties of the proposed estimating procedure, and show that with appropriately chosen threshold λ, we can ensure that the estimate produced by Algorithm 2 is minimax optimal under suitable conditions. Before proceeding, we need a couple of probabilistic bounds regarding the quality of $\widehat{\mathbf{T}}^{\mathrm{init}}$ and $\widehat{\mathbf{N}}_j$, respectively.

3. Preliminary bounds.

It is clear that the success of our estimating procedure hinges upon how close $\widehat{\mathbf{T}}^{\mathrm{init}}$ is to $\mathbf{T}$, and $\widehat{\mathbf{N}}_j$ to $\mathbf{N}_j$. We shall begin by establishing spectral bounds on them, which might also be of independent interest.

3.1. *Bounding the spectral norm of* $\widehat{\mathbf{T}}^{\mathrm{init}} - \mathbf{T}$.

We first consider bounding the spectral norm of $\widehat{\mathbf{T}}^{\mathrm{init}} - \mathbf{T}$. Write, for two tensors $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$,

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{\omega \in [d_1] \times \cdots \times [d_k]} A(\omega) B(\omega)$$

as their inner product. The spectral norm of a tensor $\mathbf{A}$ is given by

$$\|\mathbf{A}\| = \max_{u_j \in \mathbb{R}^{d_j} : \|u_1\|_{\ell_2}, \ldots, \|u_k\|_{\ell_2} \le 1} \langle \mathbf{A}, u_1 \otimes \cdots \otimes u_k \rangle.$$

The following theorem provides a probabilistic bound on the spectral norm of the difference $\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}$.

THEOREM 1. *Assume that ξ is sub-Gaussian in that there exits a $\sigma_\xi > 0$ such that for all $s \in \mathbb{R}$,*

$$\mathbb{E}\left(\exp\{s\xi\}\right) \le \exp\left(s^2\sigma_\xi^2/2\right).$$

There exists a numerical constant $C > 0$ such that, for any $\alpha \ge 1$,

$$\begin{aligned}&\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}\| \\ &\quad \le Ck^{k+3}\alpha\left(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\right)\max\left\{\sqrt{\frac{kd_{\max}d_1\cdots d_k}{n}}, \frac{kd_1\cdots d_k}{n}\right\}\log^{k+2} d_{\max},\end{aligned}$$

with probability at least $1 - d_{\max}^{-\alpha}$.

In the matrix case, that is, $k = 2$, the bound given by Theorem 1 is essentially the same as those from [16]. More importantly, Theorem 1 also highlights a key difference between matrices ($k = 2$) and higher order tensors ($k \ge 3$). To fix ideas, consider, for example, the case when $\|\mathbf{T}\|_{\ell_\infty}, \sigma_\xi \asymp (d_1\cdots d_k)^{-1/2}$. Theorem 1 implies that

$$\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}\| = O_p\left(\max\left\{\sqrt{\frac{kd_{\max}}{n}}, \frac{k(d_1\cdots d_k)^{1/2}}{n}\right\}\text{polylog}(d_{\max})\right), \tag{3.1}$$

where polylog$(\cdot)$ is a certain polynomial of the logarithmic function.

REMARK 1. The nature of the two terms on the right-hand side of (3.1) is similar to the classical Bernstein inequality and they are optimal up to k and polylog$(d_{\max})$. Indeed, by the definition of $\|\cdot\|$, we immediately have (since the tensor operator norm is lower bounded by the magnitude of any entry and the ℓ_2-norm of any fiber)

$$\begin{aligned}\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}\| \ge \max\Big\{&\max_{j\in[k]}\max_{\substack{i_{j'}\in[d_{j'}]\\ j'\neq j}}\left\|\left(\widehat{T}^{\text{init}} - T\right)(i_1,\dots,i_{j-1},:,i_{j+1},\dots,i_k)\right\|_{\ell_2}, \\ &\max_{\omega\in[d_1]\times\cdots\times[d_k]}\left|\left(\widehat{T}^{\text{init}} - T\right)(\omega)\right|\Big\}.\end{aligned} \tag{3.2}$$

Clearly, the expectations of the first and second term on the right-hand side of (3.2) are lower bounded by $\sqrt{d_{\max}/n}$ and $(d_1\cdots d_k)^{1/2}/n$, respectively.

In the matrix case, the first term on the right-hand side of (3.1) is indeed the dominating term, the very reason why the best low rank approximation of $\widehat{\mathbf{T}}^{\text{init}}$ is a good estimate of $\mathbf{T}$. For higher order tensors, however, the two different rates of convergence emerge depending on the sample size. The first term is the leading term only if

$$n \gg d_{\max}^{-1}(d_1\cdots d_k).$$

Yet, for smaller sample sizes, the second term dominates so that

$$\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}\| = O_p\left(\frac{(d_1\cdots d_k)^{1/2}\,\text{polylog}(d_{\max})}{n}\right).$$

In particular, $\widehat{\mathbf{T}}^{\text{init}}$ is consistent in terms of spectral norm if

$$n \gg \max\{d_{\max}, (d_1\cdots d_k)^{1/2}\}\cdot\text{polylog}(d_{\max}).$$

In a way, this is why we need the sample size requirement such as (1.6). It is in place to ensure that $\widehat{\mathbf{T}}^{\text{init}}$ is an consistent estimate of $\mathbf{T}$ in the sense of tensor spectral norm.

3.2. *Bounding the spectral norm of* $\widehat{\mathbf{N}}_j - \mathbf{N}_j$. Now consider bounding $\|\widehat{\mathbf{N}}_j - \mathbf{N}_j\|$. To this end, write

$$\Lambda_{\min}(\mathbf{A}) = \min\{\sigma_{\min}(\mathcal{M}_1(\mathbf{A})), \dots, \sigma_{\min}(\mathcal{M}_k(\mathbf{A}))\}$$

and

$$\Lambda_{\max}(\mathbf{A}) = \max\{\sigma_{\max}(\mathcal{M}_1(\mathbf{A})), \dots, \sigma_{\max}(\mathcal{M}_k(\mathbf{A}))\},$$

where $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and largest nonzero singular values of a matrix. Then the condition number of $\mathbf{A}$ is defined as

$$\kappa(\mathbf{A}) := \frac{\Lambda_{\max}(\mathbf{A})}{\Lambda_{\min}(\mathbf{A})}.$$

Recall that

$$\mathbf{N}_j := \mathcal{M}_j(\mathbf{T})\mathcal{M}_j(\mathbf{T})^\top$$

and

$$\widehat{\mathbf{N}}_j := \frac{(d_1 \cdots d_k)^2}{n(n-1)} \sum_{1 \le i \ne i' \le n} Y_i Y_{i'} \mathcal{M}_j(\mathbf{e}_{\omega_i})\mathcal{M}_j(\mathbf{e}_{\omega_{i'}})^\top.$$

It was proved in [28] that when $k = 3$ and $\sigma_\xi = 0$, $\widehat{\mathbf{N}}_j$ is a good estimate of $\mathbf{N}_j$. Our next result shows that this is also true for more general situations.

THEOREM 2. *There exist absolute constants* $C_1, C_2 > 0$ *such that*, *for any* $\alpha \ge 1$, *if*

$$n \ge C_1\alpha\left(\sqrt{d_1 \cdots d_k}\log d_{\max} + d_{\max}\log^2 d_{\max}\right),$$

then, *with probability at least* $1 - d_{\max}^{-\alpha}$,

$$\begin{aligned}
&\|\widehat{\mathbf{N}}_j - \mathbf{N}_j\| \\
&\quad\le C_2\biggl(\bigl(\sigma_\xi + \|\mathbf{T}\|_{\ell_\infty}\bigr)\|\mathcal{M}_j(\mathbf{T})\|\sqrt{\frac{\alpha k d_j d_1 \cdots d_k \log d_{\max}}{n}} \\
&\qquad + \alpha^3\bigl(\|\mathbf{T}\|_{\ell_\infty}^2 + \sigma_\xi^2\log^2 d_{\max}\bigr)\frac{(kd_1 \cdots d_k)^{3/2}\log^3 d_{\max}}{n}\bigl(1 + \sqrt{d_j^2/(d_1 \cdots d_k)}\bigr)\biggr).
\end{aligned}$$

Let U_j and $\widehat{U}_j$ be the top-r_j left singular vectors of $\mathbf{N}_j$ and $\widehat{\mathbf{N}}_j$, respectively. Applying Wedin's $\sin\Theta$ theorem [27], Theorem 2 immediately implies that

$$\begin{aligned}
&\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\| \\
&\quad\le C_2(\sigma_\xi + \|\mathbf{T}\|_{\ell_\infty})\frac{\|\mathcal{M}_j(\mathbf{T})\|}{\sigma_{\min}^2(\mathcal{M}_j(\mathbf{T}))}\sqrt{\frac{\alpha k d_j d_1 \cdots d_k \log d_{\max}}{n}} \\
&\qquad + C_2\alpha^3\frac{(\|\mathbf{T}\|_{\ell_\infty}^2 + \sigma_\xi^2\log^2 d_{\max})}{\sigma_{\min}^2(\mathcal{M}_j(\mathbf{T}))}\frac{(kd_1 \cdots d_k)^{3/2}\log^3 d_{\max}}{n}\left(1 + \sqrt{\frac{d_j^2}{d_1 \cdots d_k}}\right).
\end{aligned}$$

In other words, $\widehat{U}_j$s are consistent estimates of U_js if

$$\begin{aligned}
n \gtrsim \Lambda_{\min}^{-2}(\mathbf{T})\max\biggl\{&\kappa(\mathbf{T})^2\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr)^2 d_{\max} d_1 \cdots d_k \log d_{\max}, \\
&\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi \log d_{\max}\bigr)^2(d_1 \cdots d_k)^{3/2}\log^3(d_{\max})\left(1 + \sqrt{\frac{d_j^2}{d_1 \cdots d_k}}\right)\biggr\}.
\end{aligned}$$

In particular, to fix ideas, if we look at the case when $d_1 = \cdots = d_k =: d$ and $\mathbf{T}$ is well behaved in that $\kappa(\mathbf{T})$ and $\Lambda_{\min}(\mathbf{T})^{-1}$ are bounded from above, then this bound can be simplified as

$$n \gtrsim (\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 d^{3k/2} \cdot \operatorname{polylog}(d).$$

This is to be contrasted with the naive HOSVD for which we have the following.

PROPOSITION 1. *Let U_j and $\widehat{U}_j^{\mathrm{HOSVD}}$ be the top r_j singular vectors of $\mathcal{M}_j(\mathbf{T})$ and $\mathcal{M}_j(\widehat{\mathbf{T}}^{\mathrm{init}})$, respectively. Then there exists a universal constant $C > 0$ such that, for any $\alpha \geq 1$, the following bound holds with probability at least $1 - d_{\max}^{-\alpha}$:*

$$\begin{aligned}
&\|\widehat{U}_j^{\mathrm{HOSVD}}(\widehat{U}_j^{\mathrm{HOSVD}})^\top - U_j U_j^\top\|_{\ell_2} \\
&\quad \leq C \frac{(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)}{\sigma_{\min}(\mathcal{M}_j(\mathbf{T}))} \\
&\qquad \times \max\left\{\sqrt{\left(d_j \vee \frac{d_1 \cdots d_k}{d_j}\right)\frac{\alpha k d_1 \cdots d_k \log(d_{\max})}{n}}, \frac{\alpha k d_1 \cdots d_k \log(d_{\max})}{n}\right\}.
\end{aligned}$$

By Proposition 1, in the case of well-conditioned cubic tensors, to ensure that $\widehat{U}_j^{\mathrm{HOSVD}}$s are consistent, we need a sample size

$$n \gtrsim \max\{(\Lambda_{\min}^{-1}(\mathbf{T})\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) d^k \log d, \Lambda_{\min}^{-2}(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 d^{2k-1} \log d\},$$

which is much more stringent than that for $\widehat{U}_j$.

4. Performance bounds for $\widehat{\mathbf{T}}$.

We are now in position to study the performance of our estimate $\widehat{\mathbf{T}}$, as the output from Algorithm 2. Our risk bound can be characterized by the incoherence of $\mathbf{T}$ which we shall first describe. Coherence of a tensor can be defined through the singular space of its flattening. Let U be a $d \times r$ matrix with orthonormal columns. Its coherence is given by as

$$\mu(U) = \frac{d}{r} \max_{1 \leq i \leq d} \|U_{i\cdot}\|_{\ell_2}^2,$$

where $U_{i\cdot}$ is the ith row vector of U. Now for a tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ such that $\mathcal{M}_j(\mathbf{A}) = U_j \Sigma_j V_j^\top$ is its thinned singular value decomposition, we can define its coherence by

$$\mu(\mathbf{A}) = \max\{\mu(U_1), \ldots, \mu(U_k)\}.$$

Coherence of a tensor can also be measured by its spikiness:

$$\beta(\mathbf{A}) := (d_1 \cdots d_k)^{1/2} \frac{\|\mathbf{A}\|_{\ell_\infty}}{\|\mathbf{A}\|_{\ell_2}}.$$

The spikiness of a tensor is closely related to its coherence.

PROPOSITION 2. *For any $\mathbf{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$,*

$$\beta(\mathbf{A}) \leq r_1^{1/2}(\mathbf{A}) \cdots r_k^{1/2}(\mathbf{A}) \mu^{k/2}(\mathbf{A}).$$

Conversely,

$$\mu(\mathbf{A}) \leq \beta^2(\mathbf{A}) \kappa^2(\mathbf{A}).$$

4.1. *General risk bound.*

We first provide a general risk bound for $\widehat{\mathbf{T}}$ when the sample size is sufficiently large.

THEOREM 3. *Assume that ξ is sub-Gaussian in that there exits a $\sigma_\xi > 0$ such that for all $s \in \mathbb{R}$,*

$$\mathbb{E}(\exp\{s\xi\}) \le \exp(s^2\sigma_\xi^2/2).$$

There exist constants $C_1, C_2, C_3, C_4 > 0$ depending on k only such that for any fixed $\alpha \ge 1$ and $\gamma \ge C_1$, if

$$n \ge C_2\alpha[\kappa(\mathbf{T})\beta(\mathbf{T})]^{2(k-1)} d_{\max}\log d_{\max},$$

then with probability at least $1 - d_{\max}^{-\alpha}$,

$$\begin{aligned}\frac{\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_p}}{(d_1\cdots d_k)^{1/p}} &\le C_3\gamma^2\alpha^{3/2}\kappa(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)\log^{k+2} d_{\max} \\ &\quad\times\Bigg(\kappa(\mathbf{T})r_{\max}(\mathbf{T})^{(k-1)/2}\sqrt{\frac{d_{\max}}{n}} + r_{\max}(\mathbf{T})^{1/2}\frac{(d_1\cdots d_k)^{1/4}}{n^{1/2}} \\ &\quad + r_{\max}(\mathbf{T})^{(k-1)/2}\frac{(d_1\cdots d_k)^{1/2}}{n}\Bigg),\end{aligned}$$

for all $1 \le p \le 2$ where $\widehat{\mathbf{T}}$ is the output from Algorithm 2 with $\text{iter}_{\max} > C_4 \log d_{\max}$, *and*

$$\begin{aligned}\lambda = \gamma\alpha^{3/2}(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)\log^{k+2} d_{\max} \times \Bigg(\kappa(\mathbf{T})r_{\max}(\mathbf{T})^{(k-2)/2}\sqrt{\frac{d_{\max}d_1\cdots d_k}{n}} \\ + \frac{(d_1\cdots d_k)^{3/4}}{n^{1/2}} + r_{\max}(\mathbf{T})^{(k-2)/2}\frac{d_1\cdots d_k}{n}\Bigg).\end{aligned}$$

We emphasize that Theorem 3 applies to any estimate produced by power iteration after an $O(\log d_{\max})$ number of iterations. In other words, it applies beyond the best low rank approximation to $\widehat{\mathbf{T}}^{\text{init}}$. We can further simplify the risk bound in Theorem 3 when the rank $r_{\max}(\mathbf{T})$ is not too big. More specifically, we have the following.

COROLLARY 1. *Under the assumptions of Theorem 3, if, in addition,*

$$n \ge r_{\max}(\mathbf{T})^{k-2}(d_1\cdots d_k)^{1/2} \quad \text{and} \quad \kappa(\mathbf{T})^2 r_{\max}(\mathbf{T})^{k-2} d_{\max} \le (d_1\cdots d_k)^{1/2},$$

then with probability at least $1 - d_{\max}^{-\alpha}$,

$$\frac{\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_p}}{(d_1\cdots d_k)^{1/p}} \le C\gamma^2\alpha^{3/2}\kappa(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) r_{\max}(\mathbf{T})^{1/2}\frac{(d_1\cdots d_k)^{1/4}}{n^{1/2}}\log^{k+2} d_{\max}$$

for all $1 \le p \le 2$, and some constant $C > 0$ depending on k only.

To gain further insights into the risk bound in Theorem 3, it is instructive to consider the case of cubic tensors, that is, $d_1 = \cdots = d_k =: d$. By Theorem 3, we have

$$\begin{aligned} &d^{-k/p}\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_p} \\ &\quad\le C\kappa(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)\log^{k+2} d \\ &\qquad\times\Bigg(\kappa(\mathbf{T})r_{\max}(\mathbf{T})^{(k-1)/2}\sqrt{\frac{d}{n}} + r_{\max}(\mathbf{T})^{1/2}\frac{d^{k/4}}{n^{1/2}} + r_{\max}(\mathbf{T})^{(k-1)/2}\frac{d^{k/2}}{n}\Bigg).\end{aligned}$$

This rate of convergence improves those obtained earlier by [4] and [22] even though their results are obtained under more restrictive conditions. Indeed, we shall now show that, if the tensor **T** is well conditioned, much sharper performance bounds can be established for our estimate.

4.2. *Minimax optimality.* The following result shows that when the sample size is sufficiently large, power iterations starting with a good initial value indeed produces an estimate with the optimal rate of convergence, within a finite number of iterations.

THEOREM 4. *Let ξ be sub-Gaussian in that there exits a $\sigma_\xi > 0$ such that for all $s \in \mathbb{R}$,*

$$\mathbb{E}(\exp\{s\xi\}) \le \exp(s^2\sigma_\xi^2/2).$$

There are constants $C_1, C_2, C_3 > 0$ depending on k only such that the following holds. Let $\check{\mathbf{T}}$ be the output from Algorithm 1 with the number of iterations

$$\mathrm{iter}_{\max} > C_1 \log d_{\max},$$

and initial value such that

$$\max_{1\le j\le k} \|U_j^{(0)}(U_j^{(0)})^\top - U_jU_j^\top\| \le \frac{1}{2}. \tag{4.1}$$

For any fixed $\alpha > 1$, if

$$\begin{aligned} n \ge C_2 \max\bigl\{&\alpha^2 r_{\max}(\mathbf{T})^{k-2}\Lambda_{\min}^{-2}(\mathbf{T})\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr)^2 d_{\max}(d_1\cdots d_k)\log^{2(k+2)} d_{\max},\\ &\alpha\bigl(\beta(\mathbf{T})\kappa(\mathbf{T})\bigr)^{2(k-1)} d_{\max}\log(d_{\max}),\\ &\alpha r_{\max}(\mathbf{T})^{(k-2)/2}\Lambda_{\min}^{-1}(\mathbf{T})\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr) d_1\cdots d_k \log^{k+2} d_{\max},\\ &\alpha\kappa(\mathbf{T})^2\Lambda_{\min}^{-2}(\mathbf{T})\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr)^2\bigl(d_{\max} \vee r_{\max}(\mathbf{T})^{k-1}\bigr) d_1\cdots d_k \log d_{\max}\bigr\}, \end{aligned} \tag{4.2}$$

then, with probability at least $1 - d_{\max}^{-\alpha}$,

$$\frac{\|\check{\mathbf{T}} - \mathbf{T}\|_{\ell_p}}{(d_1\cdots d_k)^{1/p}} \le C_3\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr)\sqrt{\frac{\alpha(r_{\max}(\mathbf{T})d_{\max} \vee r_{\max}(\mathbf{T})^k)\log(d_{\max})}{n}},$$

for all $1 \le p \le 2$.

As an immediate consequence of Theorems 2 and 4, we have the following.

COROLLARY 2. *Suppose that ξ is sub-Gaussian as in Theorem 4. There exist constants $C_1, C_2, C_3, C_4 > 0$ depending on k only such that the following holds for any $\alpha > 1$ and $\gamma \ge C_1$. Assume that*

$$\mathrm{iter}_{\max} > C_2 \log d_{\max},$$

and

$$\begin{aligned} \lambda = \gamma\alpha^{3/2}\bigl(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\bigr)\log^{k+2} d_{\max} \times \Biggl(&\kappa(\mathbf{T})r_{\max}(\mathbf{T})^{(k-2)/2}\sqrt{\frac{d_{\max}d_1\cdots d_k}{n}}\\ &+ \frac{(d_1\cdots d_k)^{3/4}}{n^{1/2}} + r_{\max}(\mathbf{T})^{(k-2)/2}\frac{d_1\cdots d_k}{n}\Biggr). \end{aligned}$$

If

$$\begin{aligned} n \ge C_3\gamma^2 \max\{&\alpha(\beta(\mathbf{T})\kappa(\mathbf{T}))^{2(k-1)} d_{\max}\log d_{\max}, \\ &\alpha^3(\kappa^2(\mathbf{T}) \vee r_{\max}(\mathbf{T})^{k-2})\Lambda_{\min}^{-2}(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 d_{\max} d_1\cdots d_k \log^{2(k+2)} d_{\max}, \\ &\alpha^3\Lambda_{\min}^{-2}(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 (d_1\cdots d_k)^{3/2}\log^{(k+2)} d_{\max}, \\ &\alpha^{3/2} r_{\max}(\mathbf{T})^{(k-2)/2}\Lambda_{\min}^{-1}(\mathbf{T})(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) d_1\cdots d_k \log^{(k+2)} d_{\max}\} \end{aligned}$$

then, *with probability at least* $1 - d_{\max}^{-\alpha}$,

$$\frac{\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_p}}{(d_1\cdots d_k)^{1/p}} \le C_4(\sigma_\xi \vee \|\mathbf{T}\|_{\ell_\infty})\sqrt{\frac{\alpha(r_{\max}(\mathbf{T})d_{\max} \vee r_{\max}(\mathbf{T})^k)\log d_{\max}}{n}}$$

for all $1 \le p \le 2$.

It is instructive to consider the case when $d_1 = \cdots = d_k = d$, and $(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) = O(\|\mathbf{T}\|_{\ell_2}(d_1\cdots d_k)^{-1/2})$, then Corollary 2 implies that

$$d^{-k/2}\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_2} = O\biggl((\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)\sqrt{\frac{(r_{\max}(\mathbf{T})d \vee r_{\max}(\mathbf{T})^k)\log(d)}{n}}\biggr),$$

given that

$$n \gg r_{\max}(\mathbf{T})^{(k-1)/2}(d_1\cdots d_k)^{1/2}\operatorname{polylog}(d).$$

In particular, when $k = 2$, this matches the optimal bounds for noisy matrix completion; see, for example, [13, 16] and references therein. Indeed, as Theorem 5 shows that the rate of convergence achieved by $\widehat{\mathbf{T}}$ is indeed minimax optimal up to the logarithmic factor.

REMARK 2. Compared with noisy matrix completion ([16]) (e.g., $k = 2$), Corollary 2 imposes additional sample size requirements w.r.t. to the tensor spikiness. On the statistical aspect, the root cause is from the two rates (3.1) in the upper bound of $\|\widehat{\mathbf{T}}^{\text{init}} - \mathbf{T}\|$. When $k = 2$, the dominating term in (3.1) already characterizes the statistically optimal rate of noisy matrix completion. Consequently (as the proof of Theorem 3), the optimal rate is achievable if $k = 2$ even when the spectral estimates are completely noninformative. However, for higher order tensors ($k \ge 3$), the dominating term in (3.1) is generally suboptimal. To guarantee statistically optimal rates when $k \ge 3$, nontrivial spectral estimates are necessary which, in turn, imposes sample size requirement w.r.t. to tensor spikiness as in Theorem 2.

Let $\mathbb{P}_{\mathbf{T}}$ denote the joint distribution of $\{(Y_i, \omega_i) : i = 1, \ldots, n\}$ with

$$Y_i = T(\omega_i) + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma_\xi^2).$$

Denote by

$$\Theta(r_0, \beta_0) := \{\mathbf{A} \in \mathbb{R}^{d_1\times\cdots\times d_k} : r_{\max}(\mathbf{A}) \le r_0; \beta(\mathbf{A}) \le \beta_0\}.$$

THEOREM 5. *Let* $\beta_0 \ge 2$. *Then there exist absolute constants* $C_1, C_2 > 0$ *such that for any* $M \ge 0$,

$$\inf_{\tilde{\mathbf{T}}} \sup_{\mathbf{T}\in\Theta(r_0,\beta_0):\|\mathbf{T}\|_{\ell_\infty}\le M} \mathbb{P}_{\mathbf{T}}\biggl\{\frac{\|\tilde{\mathbf{T}} - \mathbf{T}\|_{\ell_p}}{(d_1\cdots d_k)^{1/p}} \ge C_1(M \wedge \sigma_\xi)\sqrt{\frac{r_0 d_{\max} \vee r_0^k}{n}}\biggr\} \ge C_2,$$

for all $1 \le p \le 2$, *where the infimum* $\tilde{\mathbf{T}}$ *is taken over all the estimators based on* $\{(Y_i, \omega_i) : 1 \le i \le n\}$.

4.3. *Random tensor model.* To better appreciate the above risk bounds, we now consider a more specific random tensor model previously studied by [22]. Let $\mathbf{T}$ be a symmetric tensor with rank r such that

$$\mathbf{T}=\sum_{i=1}^{r} u_i \otimes \cdots \otimes u_i,$$

where u_is are independent and identically distributed sub-Gaussian random vector in $\mathbb{R}^d$ with mean 0 and $\mathbb{E}(u_i \otimes u_i)=I_d$. It is not hard to see that

$$\|\mathcal{M}_j(\mathbf{T})\| \asymp_p d^{k/2}.$$

See, for example, [22]. Here, $\asymp_p$ means $\asymp$ with high probability. Meanwhile, it is clear that

$$\|\mathbf{T}\|_{\ell_2}^2=\sum_{i=1}^{r}\big(\|u_i\|_{\ell_2}^2\big)^k+2\sum_{1\le i<i'\le r}\langle u_i,u_{i'}\rangle^k \asymp_p rd^k,$$

so that

$$\sigma_{\min}(\mathcal{M}_j(\mathbf{T}) \asymp_p d^{k/2}.$$

Therefore,

$$\Lambda_{\min} \asymp_p d^{k/2} \quad \text{and} \quad \kappa(\mathbf{T}) \asymp_p 1.$$

Moreover, we have $\|\mathbf{T}\|_{\ell_\infty} \asymp_p r^{1/2}\log^{k/2} d$. If $\sigma_\xi = O(1)$, then Corollary 2 implies that, by taking

$$\lambda=\gamma\bigg(r^{(k-1)/2}\sqrt{\frac{d^{k+1}}{n}}+r^{1/2}\frac{d^{3k/4}}{n}+r^{(k-1)/2}\frac{d^k}{n}\bigg)\operatorname{polylog}(d),$$

we get

$$d^{-k/p}\|\widehat{\mathbf{T}}-\mathbf{T}\|_{\ell_p}=O_p\bigg(\sqrt{\frac{dr\log^2 d}{n}}\bigg)$$

if

$$n\ge C\gamma^2\big(dr^{k-1}+r^{(k-1)/2}d^{k/2}\big)\operatorname{polylog}(d),$$

for some absolute constant $C>0$.

5. Numerical experiments.

To complement our theoretical development, we present in this section results from several sets of numerical experiments. We begin with simulated third-order tensors where the underlying tensor $\mathbf{T}$ is generated from the following random tensor model:

$$\mathbf{T}=\sum_{k=1}^{r}\lambda(u_k\otimes v_k\otimes w_k)\in\mathbb{R}^{d\times d\times d}$$

with $\lambda=d^{3/2}$ and $U=[u_1;\ldots;u_r]\in\mathbb{R}^{d\times r}$ (also V, W) being randomly generated orthonormal columns from the eigenspace of a standard Gaussian random matrix. It is well known that $\|\lambda(u_k\otimes v_k\otimes w_k)\|_{\ell_\infty}=O(\log^{3/2} d)$ with high probability under such construction. In addition to our proposed estimator, we shall also consider the following estimator:

$$\widehat{\mathbf{T}}^{(0)}=\widehat{\mathbf{T}}^{\text{init}}\times_1\mathbf{P}_{U^{(0)}}\times_2\mathbf{P}_{V^{(0)}}\times_3\mathbf{P}_{W^{(0)}},$$

where $U^{(0)}$ ($V^{(0)}$, $W^{(0)}$, resp.) denotes the spectral initialization from U-statistics in (2.3). Note that the proposed estimator after the power iteration is given by

$$\widehat{\mathbf{T}}^{(\mathrm{iter}_{\max})} = \widehat{\mathbf{T}}^{\mathrm{init}} \times_1 \mathbf{P}_{U^{(\mathrm{iter}_{\max})}} \times_2 \mathbf{P}_{V^{(\mathrm{iter}_{\max})}} \times_3 \mathbf{P}_{W^{(\mathrm{iter}_{\max})}},$$

where $U^{(\mathrm{iter}_{\max})}$ ($V^{(\mathrm{iter}_{\max})}$, $W^{(\mathrm{iter}_{\max})}$, resp.) denote the refined estimation after $\mathrm{iter}_{\max} = 10$ power iterations. By including the estimator without power iteration, we can better appreciate the quality of spectral initialization and the effect of power iteration.

To further appreciate the merits of our approach, we also included a HOSVD based estimator:

$$\widehat{\mathbf{T}}^{(\mathrm{HOSVD})} = \widehat{\mathbf{T}}^{\mathrm{init}} \times_1 \mathbf{P}_{\widehat{U}^{\mathrm{HOSVD}}} \times_2 \mathbf{P}_{\widehat{V}^{\mathrm{HOSVD}}} \times_3 \mathbf{P}_{\widehat{W}^{\mathrm{HOSVD}}}$$

as well as the estimate proposed by [22]. We note that even though [22] considered only the noiseless case ($\sigma_\xi = 0$), their estimator can nonetheless be applied to the noisy situations.

In our simulations, we set the sample size $n = rd^\alpha$ with various choices of $\alpha \in [0, 3]$ and each observed entry is perturbed with i.i.d. Gaussian noise $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$. We set $d = 50, 100$, $r = 5$ and $\sigma_\xi = 0.2$. For each d, r, n, all four estimates were evaluated based upon 30 random realizations and the average error in estimating $\mathbf{T}$:

$$\varepsilon(\widehat{\mathbf{T}}) := \|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_2} / \|\mathbf{T}\|_{\ell_2},$$

and in estimating U

$$\varepsilon(\widehat{U}) := \|\widehat{U}\widehat{U}^\top - UU^\top\|$$

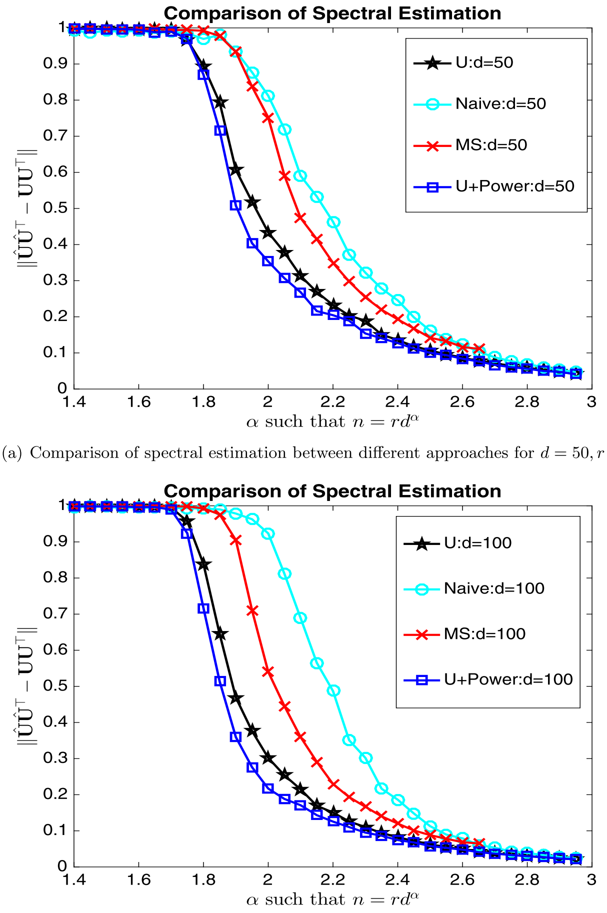
are recorded. The results are summarized by Figures 1 and 2, where "Naive" represents HOSVD based estimator; "MS" stands for the estimator from [22]; "U" corresponds to $\widehat{\mathbf{T}}^{(0)}$ and "U+Power" our proposed estimator. The plots clearly show that $\widehat{\mathbf{T}}^{(\mathrm{HOSVD})}$ requires a much larger sample size than the other estimates. It also suggests that $\widehat{\mathbf{T}}^{(0)}$ is more accurate than $\widehat{\mathbf{T}}^{\mathrm{MS}}$. Moreover, it shows that power iterations significantly improves the spectral estimation especially for larger d. Note that $\widehat{\mathbf{T}}^{\mathrm{MS}}$ can only be applied to $n \le d^3$.

Next, we apply our method to a simulated MRI brain image dataset to show the merits of our methods for denoising. The dataset can be accessed from McGill University Neurology Institute.[1] See [7] and [18] for further details. We selected "T1" modality, "1mm" slice thickness, "1%" noise, "RF" 40% and obtained therefore a third-order tensor with size 217 × 181 × 181, where each slice represents a 217 × 181 brain image. The original tensor has full rank and we project it to a tensor with multilinear ranks (20, 20, 20). In our simulations, we sampled 5%, 10%, ..., 100% entries of $\mathbf{T}$ and added i.i.d. Gaussian noise on each entries obeying distribution $\mathcal{N}(0, \sigma_\xi^2)$ where

$$\sigma_\xi = \gamma \cdot \left(\frac{\|\mathbf{T}\|_{\ell_2}^2}{217 \times 181 \times 181}\right)^{1/2}$$

with noise level $\gamma = 0.05, 0.10, 0.15 \ldots, 1.0$. We applied our reconstruction scheme to each simulated dataset and recorded the relative error (RE): $\varepsilon(\widehat{\mathbf{T}}) = \|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_2} / \|\mathbf{T}\|_{\ell_2}$. The results are presented in Figure 3 and Figure 4. It again shows that our algorithm is quite stable to noise.

[1] http://brainweb.bic.mni.mcgill.ca/brainweb/



(a) Comparison of spectral estimation between different approaches for $d = 50, r = 5$.

(b) Comparison of spectral estimation between different approaches for $d = 100, r = 5$.

FIG. 1. *Comparison of spectral estimation among four different approaches. Note that "MS" method of* [22] *only applies to* $n \le d^3$.

6. Proofs. In this section, we present the detailed proof of Theorem 1. The proofs of Theorems 2–5 and additional lemmas are collected in the Supplementary Material. We shall make use the Orlicz ψ_α-norms ($\alpha \ge 1$) of a random variable X defined as

$$\|X\|_{\psi_\alpha} := \inf\{u > 0 : \mathbb{E}\exp(|X|^\alpha/u^\alpha) \le 2\}.$$

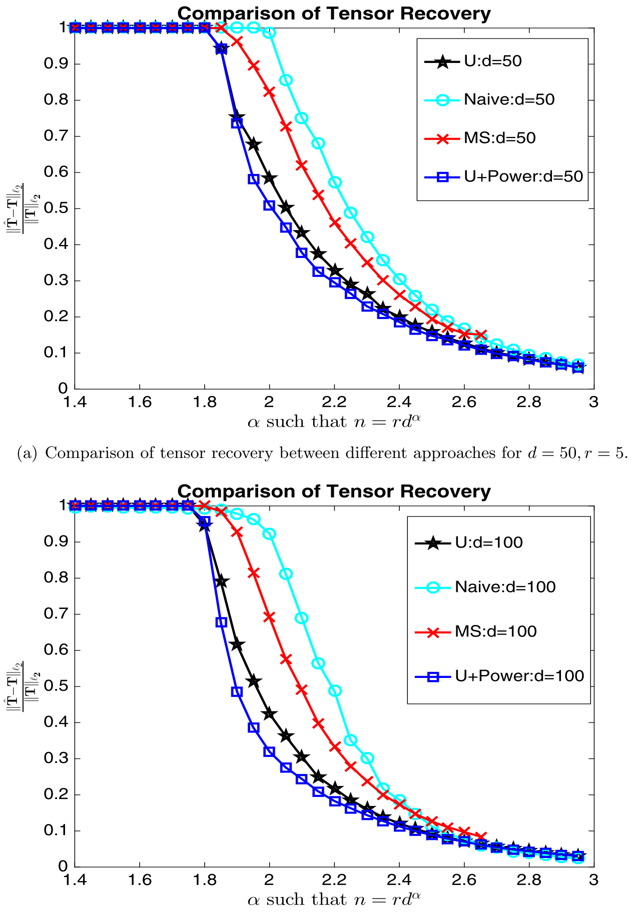
With this notion, the assumption that ξ is sub-Gaussian amounts to assuming that $\|\xi\|_{\psi_2} < +\infty$. A simple property of Orlicz norms that we shall use repeated without mentioning is the following: there exists a numerical constant $C > 0$ such that for any random variables X and Y, $\|XY\|_{\psi_1} \le C\|X\|_{\psi_2}\|Y\|_{\psi_2}$ because

$$\mathbb{E}\exp\left(\frac{|XY|}{ab}\right) \le \mathbb{E}\exp\left(\frac{X^2}{2a^2}\right)\exp\left(\frac{Y^2}{2b^2}\right) \le \mathbb{E}^{1/2}\exp\left(\frac{X^2}{a^2}\right)\mathbb{E}^{1/2}\exp\left(\frac{Y^2}{b^2}\right). \tag{6.1}$$

6.1. *Proof of Theorem* 1. The main architect of the proof follows a strategy developed by [32] for treating third-order tensors.

Symmetrization and thinning. Let $\{\varepsilon_i\}_{i=1}^n$ denote i.i.d. Rademacher random variables independent with $\{(Y_i, \mathbf{e}_{\omega_i})\}_{i=1}^n$. Define

$$\boldsymbol{\Delta} := \frac{d_1 \cdots d_k}{n}\sum_{i=1}^n \varepsilon_i Y_i \mathbf{e}_{\omega_i}.$$



(a) Comparison of tensor recovery between different approaches for $d = 50, r = 5$.

(b) Comparison of tensor recovery between different approaches for $d = 100, r = 5$.

FIG. 2. *Comparison of tensor recovery among different approaches. Note that "MS" method of* [22] *only applies to* $n \le d^3$.

We begin with symmetrization (see, e.g., [32]) and obtain for any $t > 0$,

$$\mathbb{P}\left(\left\|\frac{d_1 \cdots d_k}{n}\sum_{i=1}^{n} Y_i \mathbf{e}_{\omega_i} - \mathbf{T}\right\| \ge t\right)$$
$$\le 4\mathbb{P}(\|\mathbf{\Delta}\| \ge 2t)$$
$$+ 4\exp\left(\frac{-C_0 t^2}{C_1 d_1 \cdots d_k(\sigma_\xi^2 + \|\mathbf{T}\|_{\ell_\infty}^2)/n + C_2 t(\sigma_\xi + \|\mathbf{T}\|_{\ell_\infty})d_1 \cdots d_k/n}\right)$$

for some universal constants $C_0, C_1, C_2 > 0$ and where we used Bernstein inequality of the sum of independent sub-Gaussian random variables. It suffices to prove the upper bound of $\mathbb{P}(\|\mathbf{\Delta}\| \ge 2t)$ for any $t > 0$.

Define

$$\mathfrak{B}_{m_j, d_j} = \{0, \pm 1, \pm 2^{-1/2}, \ldots, \pm 2^{-m_j/2}\}^{d_j} \cap \{u \in \mathbb{R}^{d_j} : \|u\|_{\ell_2} \le 1\},$$

where $m_j = 2\lceil \log_2 d_j \rceil$, $j = 1, \ldots, k$. As shown by [32],

$$\|\mathbf{\Delta}\| = \sup_{\|u_j\|_{\ell_2} \le 1, j=1,\ldots,k} \langle \mathbf{\Delta}, u_1 \otimes \cdots \otimes u_k \rangle \le 2^k \sup_{u_j \in \mathfrak{B}_{m_j,d_j}} \langle \mathbf{\Delta}, u_1 \otimes \cdots \otimes u_k \rangle.$$

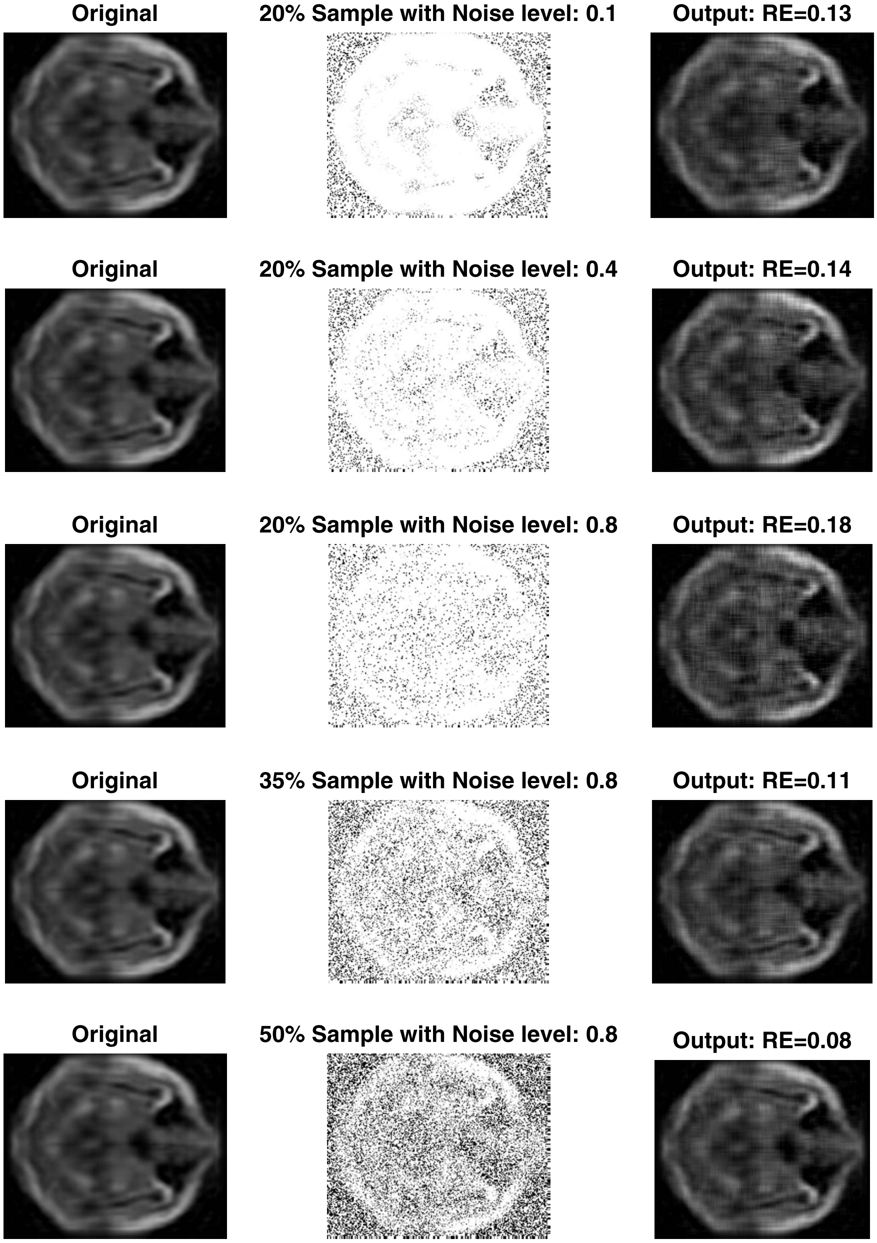

Original
20% Sample with Noise level: 0.1
Output: RE=0.13
Original
20% Sample with Noise level: 0.4
Output: RE=0.14
Original
20% Sample with Noise level: 0.8
Output: RE=0.18
Original
35% Sample with Noise level: 0.8
Output: RE=0.11
Original
50% Sample with Noise level: 0.8
Output: RE=0.08

FIG. 3. *Denoising of MRI brain image tensor. Each image is represents one slice of a tensor. The original tensor has size* $217 \times 181 \times 181$ *with multilinear ranks* $(20, 20, 20)$. *The third column represents the output of our algorithm with relative error* (*RE*) *measured as* $\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\ell_2}/\|\mathbf{T}\|_{\ell_2}$.

In fact, we can take the supreme over an even smaller set on the right-hand side.

To this end, let $\mathbf{D}_s$ be the operator that zeroes out the entries of tensor $\mathbf{A}$ whose absolute value is not $2^{-s/2}$, that is,

$$\mathbf{D}_s(\mathbf{A}) = \sum_{a_1,\dots,a_k} \mathbf{1}\{|\langle \mathbf{A}, e_{a_1} \otimes \cdots \otimes e_{a_k}\rangle| = 2^{-s/2}\}\langle \mathbf{A}, e_{a_1} \otimes \cdots \otimes e_{a_k}\rangle e_{a_1} \otimes \cdots \otimes e_{a_k},$$

where, with slight abuse on the notation, we denote by $\{e_{a_j} : 1 \le a_j \le d_j\}$ the canonical basis vectors in $\mathbb{R}^{d_j}$ for $j = 1, \dots, k$. An essential observation is that the aspect ratio of the set

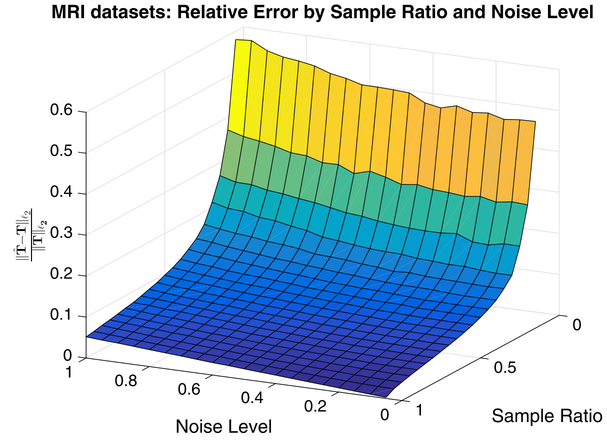


FIG. 4. *Denoising of MRI brain image tensor. The dependence of relative error on the noise level and sample ratio. We observe that our algorithm is very stable to noise level.*

$\Omega = \{\omega_i : 1 \le i \le n\}$ is typically small. More specifically, write

$$\nu_\Omega := \max_{\ell=1,\dots,k} \max_{a_j \in [d_j]: j \in [k]\setminus \ell} \bigl|\{a_\ell : (a_1, \dots, a_k) \in \Omega\}\bigr|.$$

It follows from Chernoff bound that there exists a constant $C > 0$ such that for all $\alpha \ge 1$,

$$\nu_\Omega \le C\alpha \max\left\{\frac{nd_{\max}}{d_1 d_2 \dots d_k}, k \log d_{\max}\right\} =: \nu, \tag{6.2}$$

with probability at least $1 - d_{\max}^{-\alpha}$; see, for example, [33]. We shall now proceed conditional on this event.

Obviously,

$$\sup_{u_j \in \mathfrak{B}_{m_j,d_j}} \langle \mathbf{\Delta}, u_1 \otimes \cdots \otimes u_k\rangle = \sup_{u_j \in \mathfrak{B}_{m_j,d_j}} \bigl\langle \mathbf{\Delta}, \mathcal{P}_\Omega(u_1 \otimes \cdots \otimes u_k)\bigr\rangle,$$

where $\mathcal{P}_\Omega$ is the operator that zeroes all entries of a tensor outside Ω. We shall now characterize $\mathcal{P}_\Omega \mathbf{D}_s(u_1 \otimes \cdots \otimes u_k)$. For fixed $u_j \in \mathfrak{B}_{m_j,d_j}$, $j = 1, \dots, k$, write $A_{b_j} = \{a : |u_j(a)| = 2^{-b_j/2}\}$. As shown in [32], there exist sets $\widetilde{A}_{s,b_j} \subset A_{b_j}$ such that

$$|\widetilde{A}_{s,b_j}|^2 \le \nu_\Omega \left(\prod_{j=1}^k |\widetilde{A}_{s,b_j}|\right),$$

$$(A_{b_1} \times \cdots \times A_{b_k}) \cap \Omega = (\widetilde{A}_{s,b_1} \times \cdots \times \widetilde{A}_{s,b_k}) \cap \Omega,$$

and

$$\begin{aligned}\widetilde{\mathbf{D}}_s(u_1 \otimes \cdots \otimes u_k) &:= \mathcal{P}_\Omega \widetilde{\mathbf{D}}_s(u_1 \otimes \cdots \otimes u_k) \\ &= \sum_{(b_1,\dots,b_k): b_1+\cdots+b_k = s} \mathcal{P}_{\widetilde{A}_{s,b_1} \times \cdots \times \widetilde{A}_{s,b_k}} \mathbf{D}_s(u_1 \otimes \cdots \otimes u_k).\end{aligned}$$

Now define

$$\begin{aligned}\mathfrak{B}^\star_{\Omega,m_\star} := \Biggl\{ &\sum_{0 \le s \le m_\star} \widetilde{\mathbf{D}}_s(u_1 \otimes \cdots \otimes u_k) \\ &+ \sum_{m_\star < s \le m^\star} \mathbf{D}_s(u_1 \otimes \cdots \otimes u_k) : u_j \in \mathfrak{B}_{m_j,d_j}, j = 1, \dots, k\Biggr\}\end{aligned}$$

for any $0 \le m_\star \le m^\star = \sum_{j=1}^k m_j$. Write

$$\mathfrak{B}^\star_{\nu,m_\star} = \bigcup_{\nu_\Omega \le \nu} \mathfrak{B}^\star_{\Omega,m_\star}.$$

Then

$$\|\mathbf{\Delta}\| \le 2^k \max_{\mathbf{Y} \in \mathfrak{B}^\star_{\nu,m_\star}} \langle \mathbf{Y}, \mathbf{\Delta} \rangle.$$

It is not hard to see that [32]

$$\log \operatorname{Card}(\mathfrak{B}^\star_{\nu,m_\star}) \le \frac{21}{4}(d_1 + d_2 + \cdots + d_k).$$

A refined characterization of the entropy of $\mathfrak{B}^\star_{\nu,m_\star}$ is also needed. To this end, define for any $0 \le q \le s \le m_\star$,

$$\mathfrak{D}_{\nu,s,q} := \{\mathbf{D}_s(\mathbf{Y}) : \mathbf{Y} \in \mathfrak{B}^\star_{\nu,m_\star}, \|\mathbf{D}_s(\mathbf{Y})\|^2_{\ell_2} \le 2^{q-s}\}.$$

Following an identical argument to Lemma 12 of [32], we have (for readers' convenience, we include its proof in the Appendix [29] for completeness.)

LEMMA 1. *Let $\nu \ge 1$. For all $0 \le q \le s \le m_\star$, the following bound holds*:

$$\log \operatorname{Card}(\mathfrak{D}_{\nu,s,q}) \le q s^k \log 2 + 2k^2 s^k \sqrt{\nu 2^q} L(\sqrt{\nu 2^q}, d_{\max} s^{k/2}),$$

where $L(x, y) = \max\{1, \log(ey/x)\}$.

We are now in position to bound $\|\mathbf{\Delta}\|$. Observe that

$$\begin{aligned}\|\mathbf{\Delta}\| &\le 2^k \max_{\mathbf{Y} \in \mathfrak{B}^\star_{\nu,m_\star}} \langle \mathbf{Y}, \mathbf{\Delta} \rangle \\ &= 2^k \max_{\mathbf{Y} \in \mathfrak{B}^\star_{\nu,m_\star}} \Bigl(\sum_{0 \le s \le m_\star} \langle \mathbf{D}_s(\mathbf{Y}), \mathbf{\Delta} \rangle + \langle \mathbf{S}_\star(\mathbf{Y}), \mathbf{\Delta} \rangle \Bigr),\end{aligned}$$

where $\mathbf{S}_\star(\mathbf{Y}) = \sum_{s > m_\star} \mathbf{D}_s(\mathbf{Y})$ and $m_\star$ is determined by

$$m_\star := \min\{x : x \ge m^\star \text{ or } 2k^2 x^k \sqrt{\nu 2^x} L(\sqrt{\nu 2^x}, d_{\max} x^{k/2}) \ge d_1 + \cdots + d_k\}.$$

Another simple fact is that $m_\star \le m^\star \lesssim k \lceil \log(d_{\max}) \rceil$.

Step 1: *Bounding* $|\langle \mathbf{D}_s(\mathbf{Y}), \mathbf{\Delta} \rangle|$. For any $\mathbf{Y} \in \mathfrak{B}^\star_{\nu,m_\star}$, we have $2^{-s} \le \|\mathbf{D}_s(\mathbf{Y})\|^2_{\mathrm{F}} \le 1$ and thus $\mathbf{D}_s(\mathbf{Y}) \in \bigcup_{q=0}^s \mathfrak{D}_{\nu,s,q}$. Denote $\mathbf{Y}_s = \mathbf{D}_s(\mathbf{Y})$. It suffices to develop an upper bound for

$$\max_{\mathbf{Y}_s \in \mathfrak{D}_{\nu,s,q} \setminus \mathfrak{D}_{\nu,s,q-1}} \langle \mathbf{Y}_s, \mathbf{\Delta} \rangle = \sum_{i=1}^n \langle \mathbf{Y}_s, \mathbf{Z}_i \rangle,$$

for all $0 \le q \le s$, where $\mathbf{Z}_i := \frac{d_1 \cdots d_k}{n} \varepsilon_i (\xi_i + T(\omega_i)) \mathbf{e}_{\omega_i}$, $\forall i \in [n]$. Observe that, for any fixed $\mathbf{Y}_s \in \mathfrak{D}_{\nu,s,q} \setminus \mathfrak{D}_{\nu,s,q-1}$,

$$\begin{aligned}\mathbb{E}\langle \mathbf{Y}_s, \mathbf{Z}_i \rangle^2 &\le 2\frac{(d_1 \cdots d_k)^2}{n^2} \mathbb{E}\langle \mathbf{T}, \mathbf{e}_{\omega_i} \rangle^2 \langle \mathbf{Y}_s, \mathbf{e}_{\omega_i} \rangle^2 + 2\frac{(d_1 \cdots d_k)^2}{n^2} \mathbb{E}\xi^2 \langle \mathbf{e}_{\omega_i}, \mathbf{Y}_s \rangle^2 \\ &\le 2\frac{d_1 \cdots d_k}{n^2} (\|\mathbf{T}\|^2_{\ell_\infty} + \sigma^2_\xi) \|\mathbf{Y}_s\|^2_{\ell_2} \le 2\frac{d_1 \cdots d_k}{n^2} (\|\mathbf{T}\|^2_{\ell_\infty} + \sigma^2_\xi) 2^{q-s},\end{aligned}$$

and

$$\begin{aligned}\|\langle \mathbf{Y}_s, \mathbf{Z}_i\rangle\|_{\psi_2} &\le \left\|\frac{d_1\cdots d_k}{n}\varepsilon_i\langle \mathbf{e}_{\omega_i}, \mathbf{T}\rangle\langle \mathbf{Y}_s, \mathbf{e}_{\omega_i}\rangle\right\|_{\psi_2} + \left\|\frac{d_1\cdots d_k}{n}\varepsilon_i\xi_i\langle \mathbf{e}_{\omega_i}, \mathbf{Y}_s\rangle\right\|_{\psi_2}\\ &\le C\frac{d_1\cdots d_k}{n}\big(\|\mathbf{T}\|_{\ell_\infty} + \sigma_\xi\big)2^{-s/2},\end{aligned}$$

for some constant $C > 0$, implying that $\langle \mathbf{Y}_s, \mathbf{Z}_i\rangle$ has a sub-Gaussian tail. By the Bernstein inequality for the sum of unbounded random variables,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^n \langle \mathbf{Y}_s, \mathbf{Z}_i\rangle\right| \ge t\right)\\ &\quad\le \exp\left(\frac{-C_0 t^2}{C_1 d_1\cdots d_k(\sigma_\xi^2 + \|\mathbf{T}\|_{\ell_\infty}^2)2^{q-s}/n + C_2 t(\sigma_\xi + \|\mathbf{T}\|_{\ell_\infty})d_1\cdots d_k 2^{-s/2}/n}\right),\end{aligned}$$

for some universal constants $C_0, C_1, C_2 > 0$. An application of the union bound yields

$$\begin{aligned}&\mathbb{P}\left(\max_{\mathbf{Y}_s\in\mathfrak{D}_{v,s,q}\setminus\mathfrak{D}_{v,s,q-1}}\left|\sum_{i=1}^n \langle \mathbf{Y}_s, \mathbf{Z}_i\rangle\right| \ge t\right)\\ &\quad\le |\mathfrak{D}_{v,s,q}|\\ &\qquad\times \exp\left(\frac{-C_0 t^2}{C_1 d_1\cdots d_k(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2 2^{q-s}/n + C_2 t(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k 2^{-s/2}/n}\right)\\ &\quad\le \exp\left(\frac{21}{4}(d_1+\cdots+d_k) - \frac{C_0 t^2}{C_1 d_1\cdots d_k(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2 2^{q-s}/n}\right)\\ &\qquad+ \exp\left(\log\mathrm{Card}(\mathfrak{D}_{v,s,q}) - 2^{s/2}\frac{C_0 t}{C_2(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k/n}\right).\end{aligned}$$

Recall that $m_\star \lesssim k\log(d_{\max})$,

$$\log\mathrm{Card}(\mathfrak{D}_{v,s,q}) \lesssim (k\log d_{\max})^{k+1} + 2k^2(k\log d_{\max})^k\sqrt{\nu 2^q}\,L\big(\sqrt{\nu 2^q}, d_{\max}s^{k/2}\big),$$

and

$$L\big(\sqrt{\nu 2^q}, d_{\max}s^{k/2}\big) \lesssim k\log d_{\max}.$$

By choosing

$$\begin{aligned}t \ge C_1\big(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi\big)\max\Bigg\{&2^{(q-s)/2}\sqrt{\frac{kd_{\max}d_1\cdots d_k}{n}}, 2^{-s/2}(k\log d_{\max})^{k+1}\frac{d_1\cdots d_k}{n},\\ &k^3(k\log d_{\max})^k\sqrt{v}2^{(q-s)/2}\frac{d_1\cdots d_k\log d_{\max}}{n}\Bigg\},\end{aligned}\tag{6.3}$$

we get

$$\begin{aligned}\mathbb{P}\left(\max_{\mathbf{Y}_s\in\mathfrak{D}_{v,s,q}\setminus\mathfrak{D}_{v,s,q-1}}\left|\sum_{i=1}^n \langle \mathbf{Y}_s, \mathbf{Z}_i\rangle\right| \ge t\right) &\le \exp\left(\frac{-C_0 t^2}{C_1 d_1\cdots d_k(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2 2^{q-s}/n}\right)\\ &\quad+ \exp\left(-2^{s/2}\frac{C_0 t}{C_2(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k/n}\right).\end{aligned}$$

By making the above bound uniform over all pairs $0 \le q \le s \le m_\star$, we obtain that

$$\begin{aligned}
&\mathbb{P}\biggl(\max_{\mathbf{Y}\in\mathfrak{B}^\star_{v,m_\star}}\biggl|\sum_{0\le s\le m_\star}\sum_{i=1}^{n}\langle \mathbf{Y}_s,\mathbf{Z}_i\rangle\biggr| \ge (m_\star+1)t\biggr)\\
&\quad\le 1-\binom{m_\star+2}{2}\exp\biggl(-\frac{C_0t^2}{C_1d_1\cdots d_k(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2 2^{q-s}/n}\biggr)\\
&\qquad-\binom{m_\star+2}{2}\exp\biggl(-2^{s/2}\frac{C_0t}{C_2(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k/n}\biggr).
\end{aligned}$$

Step 2: *Bounding* $\max_{\mathbf{Y}\in\mathfrak{B}^\star_{v,m_\star}}|\sum_{i=1}^{n}\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{Z}_i\rangle|$. Observe that

$$\begin{aligned}
\mathbb{E}\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{Z}_i\rangle^2 &\le 2\frac{(d_1\cdots d_k)^2}{n^2}\mathbb{E}\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{e}_{\omega_i}\rangle^2\langle\mathbf{T},\mathbf{e}_{\omega_i}\rangle^2\\
&\quad+2\frac{(d_1\cdots d_k)^2}{n^2}\mathbb{E}\xi_i^2\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{e}_{\omega_i}\rangle^2\\
&\le 2\frac{d_1\cdots d_k}{n^2}\|\mathbf{S}_\star(\mathbf{Y})\|_{\mathrm{F}}^2(\|\mathbf{T}\|_{\ell_\infty}^2+\sigma_\xi^2)\\
&\le 2^{-m_\star+1}\frac{d_1\cdots d_k}{n^2}(\|\mathbf{T}\|_{\ell_\infty}^2+\sigma_\xi^2),
\end{aligned}$$

and

$$\begin{aligned}
\|\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{Z}_i\rangle\|_{\psi_2} &\le \biggl\|\frac{d_1\cdots d_k}{n}\varepsilon_i\langle\mathbf{S}_\star,\mathbf{e}_{\omega_i}\rangle\langle\mathbf{T},\mathbf{e}_{\omega_i}\rangle\biggr\|_{\psi_2}\\
&\quad+\biggl\|\frac{d_1\cdots d_k}{n}\varepsilon_i\xi_i\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{e}_{\omega_i}\rangle\biggr\|_{\psi_2}\\
&\le C\frac{d_1\cdots d_k}{n}2^{-m_\star/2}(\|\mathbf{T}\|_{\ell_\infty}+\sigma_\xi),
\end{aligned}$$

for some constant $C>0$. Again, by the Bernstein inequality and the union bound,

$$\begin{aligned}
&\mathbb{P}\biggl(\biggl|\max_{\mathbf{Y}\in\mathfrak{B}^\star_{v,m_\star}}\sum_{i=1}^{n}\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{Z}_i\rangle\biggr|\ge t\biggr)\\
&\quad\le \exp(21(d_1+\cdots+d_k)/4)\\
&\quad\quad\times\exp\biggl(-\frac{C_0t^2}{C_1d_1\cdots d_k2^{-m_\star+1}(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2/n+C_2t(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k2^{-m_\star/2}/n}\biggr).
\end{aligned}$$

By choosing

$$t\ge C(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)\max\biggl\{2^{-(m_\star-1)/2}\sqrt{\frac{kd_{\max}d_1\cdots d_k}{n}},2^{-m_\star/2}\frac{kd_{\max}d_1\cdots d_k}{n}\biggr\},$$

we get

$$\begin{aligned}
&\mathbb{P}\biggl(\biggl|\max_{\mathbf{Y}\in\mathfrak{B}^\star_{v,m_\star}}\sum_{i=1}^{n}\langle\mathbf{S}_\star(\mathbf{Y}),\mathbf{Z}_i\rangle\biggr|\ge t\biggr)\\
&\quad\le\exp\biggl(-\frac{C_0t^2}{C_1d_1\cdots d_k2^{-m_\star+1}(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2/n}\biggr)\\
&\qquad+\exp\biggl(-\frac{C_0t}{C_2(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k2^{-m_\star/2}/n}\biggr).
\end{aligned}$$

Step 3: *Putting them together.* Combining above bounds, we conclude that if

$$t \geq C_1\big(\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi\big) \max\bigg\{\sqrt{\frac{kd_{\max}d_1\cdots d_k}{n}}, (k\log d_{\max})^{k+1}\frac{d_1\cdots d_k}{n}, \\ k^3(k\log d_{\max})^k\sqrt{v}\frac{d_1\cdots d_k \log d_{\max}}{n}, 2^{-m_\star/2}\frac{kd_{\max}d_1\cdots d_k}{n}\bigg\},$$

then

$$\mathbb{P}(\|\boldsymbol{\Delta}\| \leq (m_\star+2)t) \geq 1 - 2\binom{m_\star+2}{2}\exp\bigg(-\frac{C_0t^2}{C_1d_1\cdots d_k(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)^2/n}\bigg) \\ - 2\binom{m_\star+2}{2}\exp\bigg(-\frac{C_0t}{C_2(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi)d_1\cdots d_k/n}\bigg).$$

By the definition of $m_\star$, we have

$$2^{-m_\star/2} \lesssim \frac{\sqrt{v}}{d_{\max}}k^{3+k}\log^{k+1} d_{\max}.$$

Therefore, with probability at least $1 - d_{\max}^{-\alpha}$ for $\alpha > 1$ (by adjusting the constant C_1),

$$\|\boldsymbol{\Delta}\| \leq C_1k^{k+3}\alpha\big(\|\mathbf{T}\|_{\ell_\infty}\vee\sigma_\xi\big) \\ \times \max\bigg\{\sqrt{\frac{kd_{\max}d_1\cdots d_k}{n}}, \frac{kd_1\cdots d_k}{n}\bigg\}\log^{k+2} d_{\max}.$$

Acknowledgments. The research of Dong Xia and Ming Yuan was supported in part by NSF Grant DMS-1803450. The research of Cun-Hui Zhang was supported in part by NSF Grants DMS-1129626, DMS-1209014 and DMS-1721495.

SUPPLEMENTARY MATERIAL

Supplement to "Statistically optimal and computationally efficient low rank tensor completion from noisy entries" (DOI: 10.1214/20-AOS1942SUPP; .pdf). Supplementary information.

REFERENCES

[1] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

[2] AUFFINGER, A. and BEN AROUS, G. (2013). Complexity of random smooth functions on the high-dimensional sphere. *Ann. Probab.* **41** 4214–4247. MR3161473 https://doi.org/10.1214/13-AOP862

[3] AUFFINGER, A., BEN AROUS, G. and ČERNÝ, J. (2013). Random matrices and complexity of spin glasses. *Comm. Pure Appl. Math.* **66** 165–201. MR2999295 https://doi.org/10.1002/cpa.21422

[4] BARAK, B. and MOITRA, A. (2016). Noisy tensor completion via the sum-of-squares hierarchy. In 29*th Annual Conference on Learning Theory* 417–445.

[5] CANDES, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

[6] CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *ICML* (3) 1040–1048.

[7] COCOSCO, C. A., KOLLOKIAN, V., KWAN, R. K.-S., PIKE, G. B. and EVANS, A. C. (1997). Brainweb: Online interface to a 3D MRI simulated brain database. In *NeuroImage* Citeseer.

[8] COHEN, S. and COLLINS, M. (2012). Tensor decomposition for fast parsing with latent-variable PCFGS. In *Advances in Neural Information Processing Systems*.

[9] DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21** 1253–1278. MR1780272 https://doi.org/10.1137/S0895479896305696

[10] DE SILVA, V. and LIM, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.* **30** 1084–1127. MR2447444 https://doi.org/10.1137/06066518X

[11] HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *J. ACM* **60** 45. MR3144915 https://doi.org/10.1145/2512329

[12] JAIN, P. and OH, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems* 1431–1439.

[13] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

[14] KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583 https://doi.org/10.3150/12-BEJ486

[15] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

[16] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

[17] KREIMER, N., STANTON, A. and SACCHI, M. D. (2013). Tensor completion based on nuclear norm minimization for 5d seismic data reconstruction. *Geophysics* **78** V273–V284.

[18] KWAN, R. K.-S., EVANS, A. C. and PIKE, G. B. (1996). An extensible MRI simulator for post-processing evaluation. In *Visualization in Biomedical Computing* 135–140. Springer, Berlin.

[19] LI, N. and LI, B. (2010). Tensor completion for on-board compression of hyperspectral images. In 17*th IEEE International Conference on Image Processing* (*ICIP*) 517–520.

[20] LIM, L.-H. and COMON, P. (2010). Multiarray signal processing: Tensor decomposition meets compressed sensing. *C. R., Méc.* **338** 311–320.

[21] LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 208–220.

[22] MONTANARI, A. and SUN, N. (2018). Spectral algorithms for tensor completion. *Comm. Pure Appl. Math.* **71** 2381–2425. MR3862094 https://doi.org/10.1002/cpa.21748

[23] NION, D. and SIDIROPOULOS, N. D. (2010). Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar. *IEEE Trans. Signal Process.* **58** 5693–5705. MR2789612 https://doi.org/10.1109/TSP.2010.2058802

[24] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342 https://doi.org/10.1214/10-AOS860

[25] SEMERCI, O., HAO, N., KILMER, M. E. and MILLER, E. L. (2014). Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Process.* **23** 1678–1693. MR3191324 https://doi.org/10.1109/TIP.2014.2305840

[26] USCHMAJEW, A. (2012). Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM J. Matrix Anal. Appl.* **33** 639–652. MR2970223 https://doi.org/10.1137/110843587

[27] WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968 https://doi.org/10.1007/bf01932678

[28] XIA, D. and YUAN, M. (2019). On polynomial time methods for exact low-rank tensor completion. *Found. Comput. Math.* **19** 1265–1313. MR4029842 https://doi.org/10.1007/s10208-018-09408-6

[29] XIA, D., YUAN, M. and ZHANG, C.-H (2021). Supplement to "Statistically optimal and computationally efficient low rank tensor completion from noisy entries." https://doi.org/10.1214/20-AOS1942SUPP

[30] XIE, W., ZHU, F., JIANG, J., LIM, E.-P. and TOPICSKETCH, K. W. (2016). Real-time bursty topic detection from Twitter. *IEEE Trans. Knowl. Data Eng.* **28** 2216–2229.

[31] XU, Y. and YIN, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* **6** 1758–1789. MR3105787 https://doi.org/10.1137/120887795

[32] YUAN, M. and ZHANG, C.-H. (2016). On tensor completion via nuclear norm minimization. *Found. Comput. Math.* **16** 1031–1068. MR3529132 https://doi.org/10.1007/s10208-015-9269-5

[33] YUAN, M. and ZHANG, C.-H. (2017). Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Trans. Inf. Theory* **63** 6753–6766. MR3707566 https://doi.org/10.1109/TIT.2017.2724549