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

By Dong Xia*, Ming Yuan† and Cun-Hui Zhang‡

Hong Kong University of Science and Technology, Columbia University and Rutgers University

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 $k$th order low rank tensor under the general $\ell_p$ ($1 \leq p \leq 2$) 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 $T \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ be a $k$th order tensor, or multi-linear array. In the noisy tensor completion problem, we are interested in recovering $T$ from observations of a subset of its entries. More specifically, our sample consists of $n$ independent copies $\{(Y_i, \omega_i) : 1 \leq i \leq n\}$ of a random pair $(Y, \omega)$ obeying

\begin{equation}
Y = T(\omega) + \xi, \tag{1.1}
\end{equation}

where $\omega$ is uniformly sampled from $[d_1] \times \cdots \times [d_k]$ where $[d] = \{1, 2, \ldots, d\}$, and independent of the measurement error $\xi$ that is assumed to be a centered subgaussian random variable. Of particular interest here is the high dimensional settings where the sample size $n$ may be much smaller than the

* The research of Dong Xia was supported in part by Hong Kong RGC Grant ECS 26302019.
† The research of 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.
ambient dimension $d_1 \cdots d_k$. In this case, it may not be possible to estimate an arbitrary $k$th 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., 20, 22, 35], signal processing [e.g., 21, 27, 18, 26], latent variable modeling [e.g., 8, 6, 1, 34], 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, e.g., [5, 14, 17, 25, 15] and references therein. In particular, as shown by [17], an estimator based on nuclear norm regularization, denoted by $\hat{T}^{\text{KLT}}$, converges to $T$ at the rate of

$$
(1.2) \quad \frac{\|\hat{T}^{\text{KLT}} - T\|_2}{(d_1 d_2)^{1/2}} = O_p \left( (\|T\|_{\ell_\infty} \vee \sigma_\xi) \sqrt{\frac{r(d_1 \vee d_2) \log(d_1 \vee d_2)}{n}} \right),
$$

where $a \vee b = \max\{a, b\}$, and $\|\cdot\|_{\ell_p}$ ($p \geq 1$) denotes the vectorized $\ell_p$ 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 [17] 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 $T$, 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. In the case of matrices, SVD offers an analytic solution to the best low rank approximation essential for high dimensional matrix estimation, thanks to the Eckart-Young theorem. In contrast, for higher order tensors, there is no similar result and best low rank approximations to a higher order tensor may not even exist. See, e.g., [11]. On the other hand, as pointed out by [12], most computational problems related to higher order tensors, includ-
NOISY TENSOR COMPLETION

...ing the simple task of evaluating tensor spectral and nuclear norms, are generally NP-hard. Consequently, popular approaches such as nuclear norm minimization, while attractive for matrices, are computationally intractable for higher order tensors. This dictates that it is essential to take computational efficiency into account in devising statistically optimal estimating procedures for $T$. In fact, recent studies on a related problem, namely exact tensor completion without noise, have already pointed to many fundamental differences between matrices and higher order tensors despite their superficial similarities. See, e.g., [36, 37, 33].

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 $\hat{T}_{BM}$, such that

$$\|\hat{T}_{BM} - T\|_1^1 = O_p\left(\frac{\|T\|_* (d_{min}d_{max}^2)^{1/4} \log^2(d_{max})}{\sqrt{n}} + \|\Xi\|_1 d_{min} d_{max} + \|\Xi\|_\infty \sqrt{n}\right),$$

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 $\Xi$ is a $d_1 \times d_2 \times d_3$ random tensor whose entries are independent copies of $\xi$. More recently, the authors of [24] considered approximating a general $k$th 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 $\hat{T}_{MS}$ obeying

$$\|\hat{T}_{MS} - T\|_2 = O_p\left(\frac{\|T\|_2 r^{1/3} d^{k/6} \log^3(d)}{n^{1/3}}\right),$$

under more restrictive assumptions, where $r$ is the rank of $T$. Clearly, the bounds (1.3) and (1.4) are statistically sub-optimal in view of the degrees of freedom for low rank tensors. 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 $\ell_p$ ($1 \leq p \leq 2$) loss. We propose a computational efficient procedure based on low rank projection of an unbiased estimate of $T$, and show that, if $T$ is well conditioned, then the estimation error of the resulting estimate, denoted by $\hat{T}$, satisfies

$$\left(\frac{1}{d_1 \cdots d_k} \|\hat{T} - T\|_{\ell_p}^p\right)^{1/p} = O_p\left(\|T\|_\infty \sigma_{\xi} \sqrt{r d_{max} \log(d_{max}) n}\right),$$

under more restrictive assumptions, where $r$ is the rank of $T$. Clearly, the bounds (1.3) and (1.4) are statistically sub-optimal in view of the degrees of freedom for low rank tensors. 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.
provided that
\begin{equation}
(1.6) \\
n \geq C \left( r^{(k-1)/2}(d_1 \cdots d_k)^{1/2} \log^{k+2}(d_{\text{max}}) + r^{k-1}d_{\text{max}} \log^{2(k+2)}(d_{\text{max}}) \right),
\end{equation}

for a suitable constant \( C > 0 \). Here \( d_{\text{max}} = d_1 \vee \cdots \vee d_k \) as before. The above result continues to hold if \( r \) represents the maximum multilinear rank of \( T \).

Note that under the typical incoherent assumptions for \( T \), \( \|T\|_* \geq \|T\|_{\ell_2} \approx (d_1 \cdots d_k)^{1/2}\|T\|_{\ell_\infty} \) as we shall discuss in further details later. Therefore \( \hat{T} \) converges to \( T \) at a much faster rate than both \( \hat{T}^{BM} \) and \( \hat{T}^{MS} \). 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, e.g., [11] and [12]. 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., 3, 2]. 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 \( \hat{T}^{BM} \) 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 by [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 \( T \) 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.
However, for higher order tensor completion, there is a gap in sample size requirement between the statistical optimal and computational efficient approaches: while a sample size $n = O(d)$ may suffice to ensure consistent recovery via MLE for a $d \times d \times d$ tensor of fixed rank, $n = O(d^{3/2})$ is needed to allow for both consistent and computationally tractable estimators.

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 $T$ perfectly. See, e.g., [13, 36, 37, 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 $\xi$. In fact, much of our analysis is devoted to carefully control the adverse effect of $\xi$. On the other hand, our interest in the noisy case is in seeking a good estimate or approximation of $T$, 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 $T \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ based on observations $\{(Y_i, \omega_i) : 1 \leq i \leq n\}$ that follow $Y_i = T(\omega_i) + \xi_i, \quad i = 1, \ldots, n,$ assuming that $T$ 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 $k$th order tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ are the $d_j$ dimensional vectors

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

that is, vectors obtained by fixing all but the $j$th indices of $A$. Let $\mathcal{M}_j(A)$ be the mode-$j$ flattening of $A$ so that $\mathcal{M}_j(A)$ 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 $A$. 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
\[
\mathcal{M}_1(\mathbf{A})(i_1, (i_2 - 1)d_3 + i_3) = A(i_1, i_2, i_3), \quad \forall i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3].
\]
Denote by $r_j(\mathbf{A})$ the rank of $\mathcal{M}_j(\mathbf{A})$. The tuple $(r_1(\mathbf{A}), \ldots, r_k(\mathbf{A}))$ is often referred to the multilinear ranks of $\mathbf{A}$. Denote by
\[
r_{\text{max}}(\mathbf{A}) := \max \{r_1(\mathbf{A}), \ldots, r_k(\mathbf{A})\}.
\]

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
\[
(u_1 \otimes \cdots \otimes u_k)(i_1, \ldots, i_k) = u_1(i_1) \cdots u_k(i_k).
\]
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 $r_{\text{max}}(\mathbf{A}) \leq R(\mathbf{A}) \leq r_1(\mathbf{A}) \cdots r_k(\mathbf{A}) / r_{\text{max}}(\mathbf{A})$ 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:
\[
\min_{\mathbf{A} \in \Theta(r_1, \ldots, r_k)} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - A(\omega_i))^2 \right\}
\]
\[
= \min_{\mathbf{A} \in \Theta(r_1, \ldots, r_k)} \left\{ \frac{1}{n} \sum_{i=1}^{n} A^2(\omega_i) - \frac{2}{n} \sum_{i=1}^{n} Y_i A(\omega_i) \right\},
\]
where,
\[
\Theta(r_1, \ldots, r_k) := \left\{ \mathbf{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k} : r_k(\mathbf{A}) \leq r_k \right\}
\]
is the collection of $k$th 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., 11]. Noting that $\omega_i$s 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 \( \|A\|_2^2 \), leading to

\[
\min_{A \in \Theta(r_1, \ldots, r_k)} \left\{ (d_1 \cdots d_k)^{-1} \|A\|_2^2 - \frac{2}{n} \sum_{i=1}^n Y_i A(\omega_i) \right\}
\]

\[
= \min_{A \in \Theta(r_1, \ldots, r_k)} \left\{ \left\| \frac{d_1 \cdots d_k}{n} \sum_{i=1}^n Y_i e_{\omega_i} - A \right\|_2^2 \right\}.
\]

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

\[
\hat{T}_{\text{init}} := \frac{d_1 \cdots d_k}{n} \sum_{i=1}^n Y_i e_{\omega_i}.
\]

Similar approach can also be applied to more general sampling schemes, and was first introduced by [17] 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 \((k \geq 3)\) tensor is NP-hard in general [see, e.g., 12]. This makes it practically less meaningful to study the properties of the exact projection of \( \hat{T}_{\text{init}} \) 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

\[
\min_{A \in \Theta(r_1, \ldots, r_k)} \left\{ \left\| \hat{T}_{\text{init}} - A \right\|_2^2 \right\}
\]

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, e.g., [16].

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

\[
A = C \times_1 U_1^T \times_2 \cdots \times_k U_k^T.
\]
Here the marginal multiplication $\times_j$ between a tensor $A$ and a matrix $B$ of conformable dimension results in a tensor whose entries are defined as

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

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

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

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

$$M_j(\hat{T}_{\text{init}} \times_{j' \neq j} U_{j'}).$$

This naturally leads to Algorithm 1 which updates $C$ and $U_j$s in an iterative fashion.

**Algorithm 1 Power Iterations**

1. **Input:** $\hat{T}_{\text{init}}$, $U_j^{(0)}$ for $j = 1, 2, \ldots, k$, and parameter $\text{iter}_{\text{max}}$.
2. **Output:** $\hat{T}$.
3. Set counter $\text{iter} = 0$.
4. while $\text{iter} < \text{iter}_{\text{max}}$ do
5.   Set $\text{iter} = \text{iter} + 1$ and $j = 0$.
6.   while $j < k$ do
7.     Set $j = j + 1$.
8.     Set $U_j^{(\text{iter})}$ to be the first $r_j$ left singular vectors of
9.     $$M_j\left(\hat{T}_{\text{init}} \times_{j' < j} U_{j'}^{(\text{iter})} \times_{j' > j} U_{j'}^{(\text{iter}-1)}\right),$$
10. end while
11. end while
12. Return $\hat{T} = \hat{T}_{\text{init}} \times_{j=1}^k U_j^{(\text{iter})} (U_j^{(\text{iter})})^\top$.

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, e.g., [30] 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 $d$s) 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 $T$.

It turns out, however, that the construction of an initial value for $U_j$’s that are both close to the truth, i.e., the leading left singular vectors of $M_j(T)$, and polynomial-time computable is a fairly challenging task. An obvious choice is to start with higher order singular value decomposition [HOSVD; see, e.g., 10], and initialize $U_j$ as the left singular vectors of $M_j(\hat{T}^{\text{init}})$. It is clear that how close such an initialization is to the truth is determined by the difference $M_j(\hat{T}^{\text{init}}) - M_j(T)$. 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 [33]. Note that the column vectors of left singular vectors of $M_j(T)$ are also the leading eigenvectors of $N_j := M_j(T)M_j(T)\top$. Therefore, we could consider estimating the eigenvectors of $N_j$ instead. Specifically, we first estimate $N_j$ by the following $U$-statistic:

\begin{equation}
\hat{N}_j := \frac{(d_1 \cdots d_k)^2}{n(n-1)} \sum_{1 \leq i \neq i' \leq n} Y_i Y_{i'} M_j(e_{\omega_i}) M_j(e_{\omega_{i'}})\top.
\end{equation}

We then initialize $U_j$ as the collection of eigenvectors of $\hat{N}_j$ with eigenvalues greater than a threshold $\lambda$ to be specified later. We shall show later that $\hat{N}_j$ concentrates around $N_j$ better than $M_j(\hat{T}^{\text{init}})$ around $M_j(T)$ and therefore allows for better initialization of $U_j$’s. 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 $\lambda$, 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 $\hat{T}^{\text{init}}$ and $\hat{N}_j$, respectively.

3. Preliminary Bounds. It is clear that the success of our estimating procedure hinges upon how close $\hat{T}^{\text{init}}$ is to $T$, and $\hat{N}_j$ to $N_j$. We shall begin
Algorithm 2 Compute Estimate of $T$ from $\{(Y_i, \omega_i) : 1 \leq i \leq n\}$

Input: Observations $\{(Y_i, \omega_i) : 1 \leq i \leq n\}$, threshold $\lambda$, and parameter $\text{iter}_{\text{max}}$.

2: Output: $\hat{T}$.
   Compute $\hat{T}^{\text{init}}$ as given by (2.1).
4: Initialize $U_j$s:
   for $j = 1, \ldots, k$ do
6: Compute $\hat{N}_j$ as given by (2.3).
8: end for
Run Algorithm 1 with inputs $\hat{T}^{\text{init}}, U_j^{(0)}$s and $\text{iter}_{\text{max}}$ to get $\hat{T}$.
10: Return $\hat{T}$.

by establishing spectral bounds on them, which might also be of independent interest.

3.1. Bounding the spectral norm of $\hat{T}^{\text{init}} - T$. We first consider bounding the spectral norm of $\hat{T}^{\text{init}} - T$. Write, for two tensors $A, B \in \mathbb{R}^{d_1 \times \ldots \times d_k}$,

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

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

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

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

**Theorem 1.** Assume that $\xi$ is subgaussian in that there exits a $\sigma_\xi > 0$ such that for all $s \in \mathbb{R}$,

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

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

$$\|\hat{T}^{\text{init}} - T\| \leq Ck^{k+3} \alpha \left(\|T\|_{\ell_\infty} \vee \sigma_\xi\right) \max\left\{\sqrt{kd_{\text{max}}d_1 \ldots d_k/n}, \frac{kd_1 \ldots d_k}{n}\right\} \log^{k+2} d_{\text{max}},$$

with probability at least $1 - d^{-\alpha}$.
In the matrix case, that is $k = 2$, the bound given by Theorem 1 is essentially the same as those from [17]. More importantly, Theorem 1 also highlights a key difference between matrices ($k = 2$) and higher order tensors ($k \geq 3$). To fix ideas, consider, for example, the case when $\|T\|_{\ell_\infty, \sigma_T} \asymp (d_1 \cdots d_k)^{-1/2}$.

Theorem 1 implies that

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

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

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, because the operator norm of a tensor is lower bounded by the magnitude of its entries and the $\ell_2$-norm of its fibers, we immediately get

$$\|\hat{T}_{\text{init}} - T\| \geq \max_{j \in [k]} \max_{i, j', i \in [d_{j'}]} \| (\hat{T}_{\text{init}} - T)(i_1, \cdots, i_1, i_j, \cdots, i_k) \|_{\ell_2},$$

$$\left(3.2\right)$$

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 $\hat{T}_{\text{init}}$ is a good estimate of $T$. For higher order tensors, however, this is no longer the case and 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

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

In particular, $\hat{T}_{\text{init}}$ 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 $\hat{T}_{\text{init}}$ is an consistent estimate of $T$ in the sense of tensor spectral norm.
3.2. Bounding the spectral norm of $\tilde{N}_j - N_j$. Now consider bounding $\|\tilde{N}_j - N_j\|$. To this end, write

$$\Lambda_{\text{min}}(A) = \min \{ \sigma_{\text{min}}(M_1(A)), \ldots, \sigma_{\text{min}}(M_k(A)) \}$$

and

$$\Lambda_{\text{max}}(A) = \max \{ \sigma_{\text{max}}(M_1(A)), \ldots, \sigma_{\text{max}}(M_k(A)) \}$$

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

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

Recall that $N_j := M_j(T)M_j(T)^\top$. and

$$\tilde{N}_j := \frac{(d_1 \cdots d_k)^2}{n(n-1)} \sum_{1 \leq i \neq i' \leq n} Y_i Y_i' M_j(e_{i'}) M_j(e_{i'})^\top.$$ 

It was proved in [33] that when $k = 3$ and $\sigma_\xi = 0$, $\tilde{N}_j$ is a good estimate of $N_j$. 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 $\alpha \geq 1$, if

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

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

$$\|\tilde{N}_j - N_j\| \leq C_2 \left( (\sigma_\xi + \|T\|_{\ell_\infty}) \|M_j(T)\| \sqrt{\frac{\alpha kd_1 \cdots d_k \log d_{\text{max}}}{n}} + \alpha^3 (\|T\|_{\ell_\infty}^2 + \sigma_\xi^2 \log^2 d_{\text{max}}) \frac{(kd_1 \cdots d_k)^{3/2} \log^3 d_{\text{max}}}{n} \left( 1 + \sqrt{d_j^2/(d_1 \cdots d_k)} \right) \right).$$

Let $U_j$ and $\tilde{U}_j$ be the top-$r_j$ left singular vectors of $N_j$ and $\tilde{N}_j$ respectively. Applying Wedin’s sin $\Theta$ theorem [32], Theorem 2 immediately implies that

$$\|\tilde{U}_j \tilde{U}_j^\top - U_j U_j^\top\| \leq C_2 (\sigma_\xi + \|T\|_{\ell_\infty}) \frac{\|M_j(T)\|}{\sigma_{\text{min}}^2(M_j(T))} \sqrt{\frac{\alpha kd_1 \cdots d_k \log d_{\text{max}}}{n}} + C_2 \alpha^3 \frac{(\|T\|_{\ell_\infty}^2 + \sigma_\xi^2 \log^2 d_{\text{max}})}{\sigma_{\text{min}}^2(M_j(T))} \frac{(kd_1 \cdots d_k)^{3/2} \log^3 d_{\text{max}}}{n} \left( 1 + \sqrt{d_j^2/(d_1 \cdots d_k)} \right).$$

In other words, \( \hat{U}_j \)'s are consistent estimates of \( U_j \)'s if

\[
n \gtrsim \Lambda_{\min}^{-2}(T) \max \left\{ \kappa(T)^2 (\|T\|_{\ell_\infty} \lor \sigma_\xi)^2 d_{\max} d_1 \ldots d_k \log d_{\max}, \right.
\]

\[
\left. (\|T\|_{\ell_\infty} \lor \sigma_\xi \log d_{\max})^2 (d_1 \ldots d_k)^{3/2} \log^3(d_{\max}) \left(1 + \sqrt{\frac{d_j^2}{d_1 \ldots d_k}}\right) \right\}.
\]

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

\[
n \gtrsim (\|T\|_{\ell_\infty} \lor \sigma_\xi)^2 d^{3k/2} \cdot \text{polylog}(d).
\]

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

**Proposition 1.** Let \( U_j \) and \( \hat{U}_j^{\text{HOSVD}} \) be the top \( r_j \) singular vectors of \( M_j(T) \) and \( M_j(\hat{T}^{\text{init}}) \) respectively. Then there exists a universal constant \( C > 0 \) such that, for any \( \alpha \geq 1 \), the following bound holds with probability at least \( 1 - d^{-\alpha} \max \),

\[
\|\hat{U}_j^{\text{HOSVD}}(\hat{U}_j^{\text{HOSVD}})^\top - U_j U_j^\top\|_{\ell_2} \leq C \frac{(\|T\|_{\ell_\infty} \lor \sigma_\xi)}{\sigma_{\min}(M_j(T))} \times

\times \max \left\{ \sqrt{\left(\frac{d_j \lor d_1 \ldots d_k}{d_j}\right) \frac{\alpha k d_1 \ldots d_k \log(d_{\max})}{n}}, \frac{\alpha k d_1 \ldots d_k \log(d_{\max})}{n} \right\}.
\]

By Proposition 1, in the case of well conditioned cubic tensors, to ensure that \( \hat{U}_j^{\text{HOSVD}} \)'s are consistent, we need a sample size

\[
n \gtrsim \max \left\{ (\Lambda_{\min}^{-1}(T) \|T\|_{\ell_\infty} \lor \sigma_\xi)^2 d \log d, \Lambda_{\min}^{-2}(T)(\|T\|_{\ell_\infty} \lor \sigma_\xi)^2 d^{2k-1} \log d \right\},
\]

which is much more stringent than that for \( \hat{U}_j \).

**4. Performance Bounds for \( \hat{T} \).** We are now in position to study the performance of our estimate \( \hat{T} \), as the output from Algorithm 2. Our risk bound can be characterized by the incoherence of \( T \) 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

\[
\mu(U) = \frac{d}{r} \max_{1 \leq i \leq d} \|U_i\|_{\ell_2}^2,
\]
where $U_i$ is the $i$th row vector of $U$. Now for a tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ such that $\mathcal{M}_j(A) = U_j \Sigma_j V_j^\top$ is its thinned singular value decomposition, we can define its coherence by

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

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

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

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

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

$$\beta(A) \leq \sqrt{r_1(A)} \cdots \sqrt{r_k(A)} \mu_k(A).$$

Conversely,

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

4.1. **General risk bound.** We first provide a general risk bound for $\hat{T}$ when the sample size is sufficiently large.

**Theorem 3.** Assume that $\xi$ is subgassian in that there exits a $\sigma_\xi > 0$ such that for all $s \in \mathbb{R}$,

$$\mathbb{E} \left( \exp \{ s\xi \} \right) \leq \exp \left( s^2 \sigma^2_\xi / 2 \right).$$

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

$$n \geq C_2 \alpha \left[ \kappa(T) \beta(T) \right]^{2(k-1)} d_{\text{max}} \log d_{\text{max}},$$

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

$$\frac{\|\hat{T} - T\|_{\ell_p}}{(d_1 \ldots d_k)^{1/p}} \leq C_3 \gamma^2 \alpha^{3/2} \kappa(T) \left( \|T\|_{\ell_\infty} \vee \sigma_\xi \right) \log^{k+2} d_{\text{max}} \times
$$

$$\times \left( \kappa(T) r_{\text{max}}(T)^{k-1/2} \sqrt{d_{\text{max}} n} + r_{\text{max}}(T)^{1/2} \frac{(d_1 \ldots d_k)^{1/4}}{n^{1/2}} +
$$

$$+ r_{\text{max}}(T)^{(k-1)/2} \frac{(d_1 \ldots d_k)^{1/2}}{n} \right),$$
for all $1 \leq p \leq 2$ where $\hat{T}$ is the output from Algorithm 2 with $\text{iter}_{\text{max}} > C_4 \log d_{\text{max}}$, and

$$\lambda = \gamma \alpha^{3/2} (\|T\|_{\ell_\infty} \vee \sigma_\xi) \log^{k+2} d_{\text{max}} \times \left( \kappa(T) r_{\text{max}}(T)^{(k-2)/2} \sqrt{\frac{d_{\text{max}} d_1 \ldots d_k}{n}} + \frac{(d_1 \ldots d_k)^{3/4}}{n^{1/2}} + r_{\text{max}}(T)^{(k-2)/2} \frac{d_1 \ldots d_k}{n} \right).$$

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

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

$$n \geq r_{\text{max}}(T)^{k-2} (d_1 \ldots d_k)^{1/2}$$

and

$$\kappa(T)^2 r_{\text{max}}(T)^{k-2} d_{\text{max}} \leq (d_1 \ldots d_k)^{1/2},$$

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

$$\frac{\|\hat{T} - T\|_{\ell_p}}{(d_1 \ldots d_k)^{1/p}} \leq C \gamma^2 \alpha^{3/2} \kappa(T) \left( \|T\|_{\ell_\infty} \vee \sigma_\xi \right) r_{\text{max}}(T)^{1/2} \left( \frac{d_1 \ldots d_k}{n^{1/2}} \right) \log^{k+2} d_{\text{max}}$$

for all $1 \leq p \leq 2$, 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 $d_1 = \ldots = d_k =: d$. By Theorem 3, we have

$$d^{-k/2} \|\hat{T} - T\|_{\ell_p} \leq C \kappa(T) (\|T\|_{\ell_\infty} \vee \sigma_\xi) \log^{k+2} d \times \left( \kappa(T) r_{\text{max}}(T)^{(k-1)/2} \sqrt{\frac{d}{n}} + r_{\text{max}}(T)^{1/2} \frac{d^{k/4}}{n^{1/2}} + r_{\text{max}}(T)^{(k-1)/2} \frac{d^{k/2}}{n} \right).$$

This rate of convergence improves those obtained earlier by [4] and [24] 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 \( \xi \) be subgaussian in that there exists a \( \sigma_\xi > 0 \) such that for all \( s \in \mathbb{R} \),
\[
\mathbb{E}(\exp\{s\xi\}) \leq \exp\left(s^2 \sigma_\xi^2 / 2\right).
\]
There are constants \( C_1, C_2, C_3 > 0 \) depending on \( k \) only such that the following holds. Let \( \mathbf{T} \) be the output from Algorithm 1 with the number of iterations
\[
\text{iter}_{\text{max}} > C_1 \log d_{\text{max}},
\]
and initial value such that
\[
\max_{1 \leq j \leq k} \|U_j^{(0)}(U_j^{(0)})^\top - U_j U_j^\top\| \leq \frac{1}{2}.
\]
For any fixed \( \alpha > 1 \), if
\[
n \geq C_2 \max \left\{ \alpha^2 r_{\text{max}}(\mathbf{T})^{k-2} \Lambda_{\text{min}}^{-2}(\mathbf{T}) (\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 d_{\text{max}}(d_1 \cdots d_k) \log^{2(k+2)} d_{\text{max}},
\right.
\]
\[
\alpha \left( \beta(\mathbf{T}) \kappa(\mathbf{T}) \right)^{2(k-1)} d_{\text{max}} \log(d_{\text{max}}),
\]
\[
\alpha r_{\text{max}}(\mathbf{T})^{(k-2)/2} \Lambda_{\text{min}}^{-1}(\mathbf{T}) (\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) d_1 \cdots d_k \log^{k+2} d_{\text{max}},
\]
\[
\alpha \kappa(\mathbf{T})^2 \Lambda_{\text{min}}^{-2}(\mathbf{T}) (\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi)^2 (d_{\text{max}} \vee r_{\text{max}}(\mathbf{T})^{k-1}) d_1 \cdots d_k \log d_{\text{max}} \}
\]
then, with probability at least \( 1 - d_{\text{max}}^{-\alpha} \),
\[
\|\hat{T} - T\|_{\ell_p} \leq C_3 (\|T\|_{\ell_\infty} \vee \sigma_\xi) \sqrt{\frac{\alpha(r_{\text{max}}(\mathbf{T}) d_{\text{max}} \vee r_{\text{max}}(\mathbf{T})^k) \log(d_{\text{max}})}{n}},
\]
for all \( 1 \leq p \leq 2 \).

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

Corollary 2. Suppose that \( \xi \) is subgaussian 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 \( \gamma \geq C_1 \). Assume that
\[
\text{iter}_{\text{max}} > C_2 \log d_{\text{max}},
\]
and
\[
\lambda = \gamma \alpha^{3/2} (\|\mathbf{T}\|_{\ell_\infty} \vee \sigma_\xi) \log^{k+2} d_{\text{max}} \times \left( \kappa(\mathbf{T}) r_{\text{max}}(\mathbf{T})^{(k-2)/2} \sqrt{\frac{d_{\text{max}} d_1 \cdots d_k}{n}} + \frac{(d_1 \cdots d_k)^{3/4}}{n^{1/2}} + r_{\text{max}}(\mathbf{T})^{(k-2)/2} d_1 \cdots d_k \right).
\]
If
\[ n \geq C_3 \gamma^2 \max \left\{ \alpha \left( \beta(T) \right) \kappa(T) \right\}^{2(k-1)} d_{\text{max}} \log d_{\text{max}}, \]
\[ \alpha^3 \left( \kappa^2(T) \lor r_{\text{max}}(T)^{k-2} \Lambda_{\text{min}}^{-2}(T) \right) \left( \|T\|_{\ell_\infty} \lor \sigma_\xi \right)^2 d_{\text{max}} d_1 \ldots d_k \log^{2(k+2)} d_{\text{max}}, \]
\[ \alpha^3 \Lambda_{\text{min}}^{-2}(T) \left( \|T\|_{\ell_\infty} \lor \sigma_\xi \right)^2 (d_1 \ldots d_k)^{3/2} \log^{(k+2)} d_{\text{max}}, \]
\[ \alpha^{3/2} r_{\text{max}}(T)^{(k-2)/2} \Lambda_{\text{min}}^{-1}(T) \left( \|T\|_{\ell_\infty} \lor \sigma_\xi \right) d_1 \ldots d_k \log (k+2) d_{\text{max}} \}

then, with probability at least \( 1 - d_{\text{max}}^{-\alpha} \),
\[
\frac{\| \hat{T} - T \|_{\ell_p}}{(d_1 \ldots d_k)^{1/p}} \leq C_4 \left( \sigma_\xi \lor \|T\|_{\ell_\infty} \right) \sqrt{\alpha \left( r_{\text{max}}(T)d_{\text{max}} \lor r_{\text{max}}(T)^k \right) \log d_{\text{max}}} \frac{d_{\text{max}}}{n}
\]
for all \( 1 \leq p \leq 2 \).

It is instructive to consider the case when \( d_1 = \ldots = d_k = d \), and \( \|T\|_{\ell_\infty} \lor \sigma_\xi = O(\|T\|_{\ell_2} (d_1 \ldots d_k)^{-1/2}) \), then Corollary 2 implies that
\[
d^{-k/2}\| \hat{T} - T \|_{\ell_2} = O \left( \left( \|T\|_{\ell_\infty} \lor \sigma_\xi \right) \sqrt{\left( r_{\text{max}}(T)d \lor r_{\text{max}}(T)^k \right) \log (d)} \right),
\]
given that
\[
n \gg r_{\text{max}}(T)^{(k-1)/2} (d_1 \ldots d_k)^{1/2} \text{polylog}(d).
\]
In particular, when \( k = 2 \), this matches the optimal bounds for noisy matrix completion. See, e.g., [14, 17] and references therein. Indeed, as Theorem 5 shows that the rate of convergence achieved by \( \hat{T} \) is indeed minimax optimal up to the logarithmic factor.

Compared with noisy matrix completion (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 in the upper bound of \( \|\hat{T}^{\text{init}} - T\| \). 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 non-informative. However, for higher order tensors (\( k \geq 3 \)), the dominating term in (3.1) is generally sub-optimal. To guarantee statistically optimal rates when \( k \geq 3 \), non-trivial spectral estimates are necessary which, in turn, imposes sample size requirement w.r.t. to tensor spikiness as in Theorem 2.
Let $\mathbb{P}_T$ denote the joint distribution of $\{(Y_i, \omega_i) : i = 1, \ldots, n\}$ with

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

Denote by

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

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

$$\inf_{\tilde{T}} \sup_{T \in \Theta(r_0, \beta_0) : \|T\|_{\ell^\infty} \leq M} \mathbb{P}_T \left\{ \|\tilde{T} - T\|_{\ell_p} \geq C_1 (M \wedge \sigma) \sqrt{\frac{r_0 d_{\max} \lor r_k}{n}} \right\} \geq C_2,$$

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

### 4.3. Random tensor model.

To better appreciate the above risk bounds, we now consider a more specific random tensor model previously studied by [24]. Let $T$ be a symmetric tensor with rank $r$ such that

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

where $u_i$s are independent and identically distributed subgaussian 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

$$\sigma_{\min}(M_j(T)) \asymp p \frac{d^k}{2}.$$

Therefore, $\mathbb{P}_T \left\{ \|M_j(T)\| \asymp_p d^k/2 \right\}$.

See, e.g., [24]. Here $\asymp_p$ means $\asymp$ with high probability. Meanwhile, it is clear that

$$\|T\|_{\ell^2_2}^2 = \sum_{i=1}^r \|u_i\|_{\ell^2_2}^2 + 2 \sum_{1 \leq i < i' \leq r} \langle u_i, u_{i'} \rangle^k \asymp_p r d^k,$$

so that

$$\sigma_{\min}(M_j(T)) \asymp_p d^{k/2}. $$

Therefore,

$$\Lambda_{\min} \asymp_p d^{k/2} \quad \text{and} \quad \kappa(T) \asymp_p 1.$$
Moreover, we have \( \|T\|_{\ell_\infty} \asymp p^{1/2} \log^{k/2} d \). If \( \sigma_\xi = O(1) \), then Corollary 2 implies that, by taking
\[
\lambda = \gamma \left( r^{(k-1)/2} \sqrt{d^{k+1}/n} + r^{1/2} d^{3k/4} / n + r^{(k-1)/2} d^{k}/n \right) \text{polylog}(d),
\]
we get
\[
d^{-k/p} \|\hat{T} - T\|_{\ell_p} = O_p \left( \sqrt{dr \log^2 d} / n \right)
\]
if
\[
n \geq C \gamma^2 \left( dr^{k-1} + r^{(k-1)/2} d^{k/2} \right) \text{polylog}(d),
\]
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 \( T \) is generated from the following random tensor model
\[
T = \sum_{k=1}^{r} \lambda(u_k \otimes v_k \otimes w_k) \in \mathbb{R}^{d \times d \times d}
\]
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:
\[
\hat{T}^{(0)} = \hat{T}^{\text{init}} \times_1 P_{U^{(0)}} \times_2 P_{V^{(0)}} \times_3 P_{W^{(0)}}.
\]
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
\[
\hat{T}^{(\text{itermax})} = \hat{T}^{\text{init}} \times_1 P_{U^{(\text{itermax})}} \times_2 P_{V^{(\text{itermax})}} \times_3 P_{W^{(\text{itermax})}}
\]
where \( U^{(\text{itermax})} \) (\( V^{(\text{itermax})}, W^{(\text{itermax})} \) resp.) denote the refined estimation after \( \text{itermax} = 10 \) 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 an HOSVD based estimator:

$$\hat{T}^{(\text{HOSVD})} = \hat{T}^{\text{init}} \times_1 P_{\hat{U}^{\text{HOSVD}}} \times_2 P_{\hat{V}^{\text{HOSVD}}} \times_3 P_{\hat{W}^{\text{HOSVD}}}.$$ 

as well as the estimate proposed by [24]. We note that even though [24] 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 $T$:

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

and in estimating $U$

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

are recorded. The results are summarized by Figures 1 and 2, where “Naive” represents HOSVD based estimator; “MS” stands for the estimator from [24]; “U” corresponds to $\hat{T}^{(0)}$ and “U+Power” our proposed estimator. The plots clearly show that $\hat{T}^{(\text{HOSVD})}$ requires a much larger sample size than the other estimates. It also suggests that $\hat{T}^{(0)}$ is more accurate than $\hat{T}^{\text{MS}}$. Moreover, it shows that power iterations significantly improves the spectral estimation especially for larger $d$. Note that $\hat{T}^{\text{MS}}$ can only be applied to $n \leq d^3$.

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*. See [7] and [19] for further details. We selected “T1” modality, “1mm” slice thickness, “1%” noise, “RF” 40% and obtained therefore a third-order tensor with size $217 \times 181 \times 181$, where each slice represents a $217 \times 181$ 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 $T$ and added i.i.d. Gaussian noise on each entries obeying distribution $\mathcal{N}(0, \sigma_\xi^2)$ where

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

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):

*http://brainweb.bic.mni.mcgill.ca/brainweb/
NOISY TENSOR COMPLETION

1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3
α such that \( n = rd^\alpha \)

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
\[ \| \hat{U} \hat{U}^\top - UU^\top \| \]

Comparison of Spectral Estimation

U:d=50
Naive:d=50
MS:d=50
U+Power:d=50

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

1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3
α such that \( n = rd^\alpha \)

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
\[ \| \hat{U} \hat{U}^\top - UU^\top \| \]

Comparison of Spectral Estimation

U:d=100
Naive:d=100
MS:d=100
U+Power:d=100

(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 [24] only applies to \( n \leq d^3 \).
(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 [24] only applies to \( n \leq d^3 \).
\( \varepsilon(\hat{T}) = \|\hat{T} - T\|_{\ell_2}/\|T\|_{\ell_2} \). The results are presented in Figure 3 and Figure 4. It again shows that our algorithm is quite stable to noise.

6. Proofs. In this section, we present the detailed proof of Theorem 1. The proofs of Theorem 2-5 and additional lemmas are collected in the supplementary file. We shall make use the Orlicz \( \psi_{\alpha} \)-norms \((\alpha \geq 1)\) of a random variable \( X \) defined as

\[
\|X\|_{\psi_\alpha} := \inf \left\{ u > 0 : \mathbb{E} \exp \left( \frac{|X|^\alpha}{u^\alpha} \right) \leq 2 \right\}.
\]

With this notion, the assumption that \( \xi \) is subgaussian amounts to assuming that \( \|\xi\|_{\psi_2} < +\infty \). A simple property of Orlicz norms that we shall use repeatedly without mentioning is the following: there exists a numerical constant \( C > 0 \) such that for any random variables \( X \) and \( Y \),

\[
\|XY\|_{\psi_1} \leq C \|X\|_{\psi_2} \|Y\|_{\psi_2}.
\]

6.1. Proof of Theorem 1. The main architect of the proof follows a strategy developed by [36] for treating third order tensors.

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

\[
\Delta := \frac{d_1 \cdots d_k}{n} \sum_{i=1}^n \varepsilon_i Y_i e_{\omega_i}.
\]

We begin with symmetrization (see, e.g., [36]) and obtain for any \( t > 0 \),

\[
\mathbb{P} \left( \left\| \frac{d_1 \cdots d_k}{n} \sum_{i=1}^n Y_i e_{\omega_i} - T \right\| \geq t \right) \leq 4 \mathbb{P}(\|\Delta\| \geq 2t) - C_0 t^2
\]

\[
+ 4 \exp \left( \frac{-C_0 t^2}{C_1 d_1 \cdots d_k (\sigma_\xi^2 + \|T\|_{\ell_\infty}^2) / n + C_2 t (\sigma_\xi + \|T\|_{\ell_\infty}) d_1 \cdots d_k / n} \right)
\]

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 \( \mathbb{P}(\|\Delta\| \geq 2t) \) for any \( t > 0 \).

Define

\[
\mathcal{A}_{m_j, d_j} = \left\{ 0, \pm 1, \pm 2^{-1/2}, \ldots, \pm 2^{-m_j/2} \right\}^{d_j} \cap \left\{ u \in \mathbb{R}^{d_j} : \|u\|_{\ell_2} \leq 1 \right\}.
\]
Fig 3. Denoising of MRI brain image tensor. Each image 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 $\|\hat{T} - T\|_2 / \|T\|_2$. 
where \( m_j = 2^\left\lfloor \log_2 d_j \right\rfloor \), \( j = 1, \ldots, k \). As shown by [36],
\[
\| \Delta \| = \sup_{\| u_j \|_2 \leq 1} \langle \Delta, u_1 \otimes \cdots \otimes u_k \rangle \leq 2^k \sup_{u_j \in \mathcal{B}_{m_j, d_j}} \langle \Delta, u_1 \otimes \cdots \otimes u_k \rangle.
\]

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

To this end, let \( D_s \) be the operator that zeroes out the entries of tensor \( A \) whose absolute value is not \( 2^{-s/2} \), that is
\[
D_s(A) = \sum_{a_1, \ldots, a_k} 1\{|\langle A, e_{a_1} \otimes \cdots \otimes e_{a_k} \rangle| = 2^{-s/2}\} \langle A, e_{a_1} \otimes \cdots \otimes e_{a_k} \rangle e_{a_1} \otimes \cdots \otimes e_{a_k},
\]
where, with slight abuse on the notations, we denote by \( \{e_{a_j} : 1 \leq a_j \leq d_j\} \) the canonical basis vectors in \( \mathbb{R}^{d_j} \), for \( j = 1, \ldots, k \). An essential observation is that the aspect ratio of the set \( \Omega = \{\omega_i : 1 \leq i \leq n\} \) is typically small. More specifically, write
\[
\nu_\Omega := \max_{\ell=1,\ldots,k} \max_{a_j \in [d_j]} \max_{j \in [k] \setminus \ell} |\{a_\ell : (a_1, \ldots, a_k) \in \Omega\}|.
\]

It follows from Chernoff bound that there exists a constant \( C > 0 \) such that for all \( \alpha \geq 1 \),
\[
(6.2) \quad \nu_\Omega \leq C \alpha \max \left\{ \frac{nd_{\max}}{d_1 d_2 \cdots d_k}, k \log d_{\max} \right\} =: \nu,
\]
with probability at least $1 - d_{\max}^\alpha$. See, e.g., [37]. We shall now proceed conditional on this event.

Obviously,

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

where $\mathcal{P}_\Omega$ is the operator that zeroes all entries of a tensor outside $\Omega$. We shall now characterize $\mathcal{P}_\Omega \mathcal{D}_s(u_1 \otimes \cdots \otimes u_k)$. For fixed $u_j \in \mathcal{B}_{m_j,d_j}, j = 1, \ldots, k$, write $A_{b_j} = \{a : |u_j(a)| = 2^{-b_j/2}\}$. As shown in [36], there exist sets $\tilde{A}_{s,b_j} \subset A_{b_j}$ such that

$$|\tilde{A}_{s,b_j}|^2 \leq \nu \Omega \left( \prod_{j=1}^k |\tilde{A}_{s,b_j}| \right),$$

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

and

$$\tilde{\mathcal{D}}_s(u_1 \otimes \cdots \otimes u_k) := \mathcal{P}_\Omega \mathcal{D}_s(u_1 \otimes \cdots \otimes u_k) = \sum_{(b_1, \ldots, b_k) : b_1 + \ldots + b_k = s} \mathcal{P}_{\tilde{A}_{s,b_1} \times \cdots \times \tilde{A}_{s,b_k}} \mathcal{D}_s(u_1 \otimes \cdots \otimes u_k).$$

Now define

$$\mathcal{B}_{\nu,m}^* := \left\{ \sum_{0 \leq s \leq m} \tilde{\mathcal{D}}_s(u_1 \otimes \cdots \otimes u_k) + \sum_{m_\nu < s \leq m^*} \mathcal{D}_s(u_1 \otimes \cdots \otimes u_k) : u_j \in \mathcal{B}_{m_j,d_j}, j = 1, \ldots, k \right\}$$

for any $0 \leq m^* = \sum_{j=1}^k m_j$. Write

$$\mathcal{B}_{\nu,m}^* = \bigcup_{\nu_\Omega \leq \nu} \mathcal{B}_{\nu_\Omega,m_\nu}^*.$$

Then

$$\|\Delta\| \leq 2^k \max_{\mathbf{Y} \in \mathcal{B}_{\nu,m}^*} \langle \mathbf{Y}, \Delta \rangle.$$

It is not hard to see that [36]

$$\log \text{Card}(\mathcal{B}_{\nu,m}^*) \leq \frac{21}{4} (d_1 + d_2 + \ldots + d_k).$$
A refined characterization of the entropy of $B_{ν,m}^*$ is also needed. To this end, define for any $0 \leq q \leq s \leq m_*$,
\[ \mathcal{D}_{v,s,q} := \left\{ D_s(Y) : Y \in B_{ν,m}^*, \|D_s(Y)\|_{\ell_2}^2 \leq 2^{q-s} \right\}. \]

Following an identical argument to Lemma 12 of [36], we have (for readers’ convenience, we include its proof in the Appendix for completeness.)

**Lemma 1.** Let $ν \geq 1$. For all $0 \leq q \leq s \leq m_*$, the following bound holds
\[ \log \text{Card}(\mathcal{D}_{v,s,q}) \leq qs^k \log 2 + 2k^2 s^k \sqrt{ν} L(\sqrt{ν}, \sigma_{max} s^{k/2}), \]
where $L(x, y) = \max \{1, \log(ey/x)\}$.

We are now in position to bound $\|\Delta\|$. Observe that
\[ \|\Delta\| \leq 2^k \max_{Y \in B_{ν,m}^*} \langle Y, \Delta \rangle \]
\[ = 2^k \max_{Y \in B_{ν,m}^*} \left( \sum_{0 \leq s \leq m_*} \langle D_s(Y), \Delta \rangle + \langle S_s(Y), \Delta \rangle \right), \]
where $S_s(Y) = \sum_{s > m_*} D_s(Y)$ and $m_*$ is determined by
\[ m_* := \min \{ x : x \geq m^* \text{ or } 2k^2 x^k \sqrt{ν} L(\sqrt{ν}, \sigma_{max} x^{k/2}) \geq d_1 + \ldots + d_k \}.
\]

Another simple fact is that $m_* \leq m^* \lesssim k \lceil \log(\sigma_{max}) \rceil$.

**Step 1: bounding $\|D_s(Y), \Delta\|$.** For any $Y \in B_{ν,m}^*$, we have $2^{-s} \leq \|D_s(Y)\|_{\ell_2} \leq 1$ and thus $D_s(Y) \in \cup_{q=0}^s \mathcal{D}_{v,s,q}$. Denote $Y_s = D_s(Y)$. It suffices to develop an upper bound for
\[ \max_{Y_s \in \mathcal{D}_{v,s,q} \setminus \mathcal{D}_{v,s,q-1}} \langle Y_s, \Delta \rangle = \sum_{i=1}^{n} \langle Y_s, Z_i \rangle, \]
for all $0 \leq q \leq s$, where $Z_i := \frac{d_1 \ldots d_k}{n} \varepsilon_i (\xi_i + T(ω_i))e_{ω_i}, \forall i \in [n]$. Observe that, for any fixed $Y_s \in \mathcal{D}_{v,s,q} \setminus \mathcal{D}_{v,s,q-1},$
\[ \mathbb{E} \langle Y_s, Z_i \rangle^2 \leq \frac{(d_1 \ldots d_k)^2}{n^2} \mathbb{E} \langle T, e_{ω_i} \rangle^2 \langle Y_s, e_{ω_i} \rangle^2 + 2 \frac{(d_1 \ldots d_k)^2}{n^2} \mathbb{E} \xi^2 \langle e_{ω_i}, Y_s \rangle^2 \]
\[ \leq 2 \frac{(d_1 \ldots d_k)^2}{n^2} (\|T\|_{\ell_\infty}^2 + \sigma_{max}^2) \|Y_s\|_{\ell_2}^2 \leq 2 \frac{(d_1 \ldots d_k)^2}{n^2} (\|T\|_{\ell_\infty}^2 + \sigma_{max}^2) 2^{q-s}, \]
and
\[
\|\langle Y_s, Z_i \rangle \|_{\psi_2} \leq \left\| \frac{d_1 \cdots d_k}{n} \varepsilon_i \langle e_{\omega_i}, T \rangle \langle Y_s, e_{\omega_i} \rangle \right\|_{\psi_2} + \left\| \frac{d_1 \cdots d_k}{n} \varepsilon_i \xi_i \langle e_{\omega_i}, Y_s \rangle \right\|_{\psi_2} \\
\leq C \frac{d_1 \cdots d_k}{n} (\|T\|_{\ell_\infty} + \sigma_\xi)^{2-s/2},
\]
for some constant \( C > 0 \), implying that \( \langle Y_s, Z_i \rangle \) has a subgaussian tail. By Bernstein inequality for sum of unbounded random variables,
\[
P\left( \left| \sum_{i=1}^n \langle Y_s, Z_i \rangle \right| \geq t \right) \leq \exp\left( -\frac{C_0 t^2}{C_1 d_1 \cdots d_k (\|T\|_{\ell_\infty} + \sigma_\xi) 2^{2q-s}/n} \right),
\]
for some universal constants \( C_0, C_1, C_2 > 0 \). An application of the union bound yields
\[
P\left( \max_{Y_s \in \mathcal{D}_{v,s,q}} \left| \sum_{i=1}^n \langle Y_s, Z_i \rangle \right| \geq t \right) \leq \left| \mathcal{D}_{v,s,q} \right| \times \exp\left( -\frac{C_0 t^2}{C_1 d_1 \cdots d_k (\|T\|_{\ell_\infty} \vee \sigma_\xi) 2^{2q-s}/n} \right) \leq \exp\left( \frac{\log \text{Card}(\mathcal{D}_{v,s,q}) - 2^{s/2} - \frac{C_0 t}{C_2 (\|T\|_{\ell_\infty} \vee \sigma_\xi) 2^{s-n}/n}}{C_0 t} \right).
\]
Recall that \( m_* \lesssim k \log (d_{\max}) \),
\[
\log \text{Card}(\mathcal{D}_{v,s,q}) \lesssim (k \log d_{\max})^{k+1} + 2k^2 (k \log d_{\max})^k \sqrt{\nu \nu} L(\sqrt{\nu \nu}, d_{\max} s^{k/2}),
\]
and
\[
L(\sqrt{\nu \nu}, d_{\max} s^{k/2}) \lesssim k \log d_{\max}.
\]
By choosing
\[
t \geq C_1 (\|T\|_{\ell_\infty} \vee \sigma_\xi) \max\left\{ \left( \frac{k \log d_{\max}^{k+1} d_1 \cdots d_k}{n} \right)^{2(q-s)/2}, \left( \frac{k \log d_{\max}^{k+1} d_1 \cdots d_k}{n} \right)^{2-s/2}, \left( \frac{k \log d_{\max}^{k+1} d_1 \cdots d_k}{n} \right)^{k} \right\},
\]
(6.3)
we get
\[
\Pr\left( \max_{Y \in \mathcal{Y}_{v,s,q}} \left| \sum_{i=1}^{n} \langle Y_s, Z_i \rangle \right| \geq t \right) \leq \exp \left( \frac{-C_0 t^2}{C_1 d_1 \ldots d_k (\|T\|_{\ell_\infty} \vee \sigma_\xi)^2 2^{q-s}/n} \right) + \exp \left( -\frac{2s/2}{C_2 (\|T\|_{\ell_\infty} \vee \sigma_\xi) d_1 \ldots d_k / n} \right).
\]

By making the above bound uniform over all pairs \(0 \leq q \leq s \leq m_*,\) we obtain that
\[
\Pr\left( \max_{Y \in \mathcal{Y}_{v,m_*}} \left| \sum_{0 \leq s \leq m_*} \sum_{i=1}^{n} \langle Y_s, Z_i \rangle \right| \geq (m_* + 1)t \right)
\leq 1 - \left( \frac{m_* + 2}{2} \right) \exp \left( -\frac{C_0 t^2}{C_1 d_1 \ldots d_k (\|T\|_{\ell_\infty} \vee \sigma_\xi)^2 2^{q-s}/n} \right) - \left( \frac{m_* + 2}{2} \right) \exp \left( -\frac{2s/2}{C_2 (\|T\|_{\ell_\infty} \vee \sigma_\xi) d_1 \ldots d_k / n} \right).
\]

**Step 2: bounding** \(\max_{Y \in \mathcal{Y}_{v,m_*}} \left| \sum_{i=1}^{n} \langle S_s(Y), Z_i \rangle \right|.\) Observe that
\[
\mathbb{E}(S_s(Y), Z_i)^2 \leq 2 (d_1 \ldots d_k)^2 \mathbb{E}(S_s(Y), e_{\omega_1})^2 \langle T, e_{\omega_1} \rangle^2 + 2 \frac{(d_1 \ldots d_k)^2}{n^2} \mathbb{E}(S_s(Y), e_{\omega_1})^2 \mathbb{E}(S_s(Y), e_{\omega_1})^2 \leq 2 \frac{(d_1 \ldots d_k)^2}{n^2} \|S_s(Y)\|_{\ell_2}^2 (\|T\|_{\ell_\infty}^2 + \sigma_\xi^2)
\leq 2^{-m_* + 1} \frac{d_1 \ldots d_k}{n^2} (\|T\|_{\ell_\infty}^2 + \sigma_\xi^2),
\]

and
\[
\|S_s(Y), Z_i\|_{\ell_2^2} \leq \frac{d_1 \ldots d_k}{n} e_i \langle S_s(Y), e_{\omega_1} \rangle \langle T, e_{\omega_1} \rangle + \frac{d_1 \ldots d_k}{n} e_i \langle S_s(Y), e_{\omega_1} \rangle \|Z\|_{\ell_2^2} \leq C \frac{d_1 \ldots d_k}{n} 2^{-m_* / 2} (\|T\|_{\ell_\infty} + \sigma_\xi),
\]

for some constant \(C > 0.\) Again, by Bernstein inequality and the union bound,
\[
\Pr\left( \max_{Y \in \mathcal{Y}_{v,m_*}} \sum_{i=1}^{n} \langle S_s(Y), Z_i \rangle \geq t \right) \leq \exp \left( 21 (d_1 + \ldots + d_k) / 4 \right) \times \exp \left( -\frac{C_0 t^2}{C_1 d_1 \ldots d_k 2^{-m_* + 1} (\|T\|_{\ell_\infty} \vee \sigma_\xi)^2 / n + C_2 t (\|T\|_{\ell_\infty} \vee \sigma_\xi) d_1 \ldots d_k 2^{-m_* / 2} / n} \right).
\]
By choosing
\[ t \geq C(\|T\|_{\ell_\infty} \vee \sigma_\xi) \max \left\{ 2^{-(m_*-1)/2} \sqrt{\frac{kd_{\max}d_1 \ldots d_k}{n}}, \ 2^{-m_*/2} \frac{kd_{\max}d_1 \ldots d_k}{n} \right\}, \]
we get
\[
\mathbb{P}\left( \max_{Y \in \mathcal{B}_{\nu, m_*}^v} \sum_{i=1}^n \langle S_*(Y), Z_i \rangle \geq t \right) 
\leq \exp\left( -\frac{C_0 t^2}{C_1 d_1 \ldots d_k 2^{-m_*+1} (\|T\|_{\ell_\infty} \vee \sigma_\xi)^2/n} \right) 
+ \exp\left( -\frac{C_0 t}{C_2 (\|T\|_{\ell_\infty} \vee \sigma_\xi) d_1 \ldots d_k 2^{-m_*/2}/n} \right).
\]

**Step 3: putting them together.** Combining above bounds, we conclude that if
\[
t \geq C_1 (\|T\|_{\ell_\infty} \vee \sigma_\xi) \max \left\{ \sqrt{\frac{kd_{\max}d_1 \ldots d_k}{n}}, \ (k \log d_{\max})^{k+1} \frac{d_1 \ldots d_k}{n}, \ k^3 (k \log d_{\max})^k \sqrt{\nu d_{\max}} \right\},
\]
then
\[
\mathbb{P}\left( \|\Delta\| \leq (m_* + 2)t \right) \geq 1 - 2^{m_* + 2} \frac{m_* + 2}{2} \exp\left( -\frac{C_0 t^2}{C_1 d_1 \ldots d_k (\|T\|_{\ell_\infty} \vee \sigma_\xi)^2/n} \right) 
- 2^{m_* + 2} \frac{m_* + 2}{2} \exp\left( -\frac{C_0 t}{C_2 (\|T\|_{\ell_\infty} \vee \sigma_\xi) d_1 \ldots d_k/n} \right).
\]

By the definition of \( m_* \), we have
\[
2^{-m_*/2} \lesssim \frac{\sqrt{\nu}}{d_{\max}} k^{3+k} \log^{k+1} d_{\max}.
\]

Therefore, with probability at least \( 1 - d_{\max}^{-\alpha} \) for \( \alpha > 1 \) (by adjusting the constant \( C_1 \)),
\[
\|\Delta\| \leq C_1 k^{k+3} \alpha (\|T\|_{\ell_\infty} \vee \sigma_\xi) \times
\max \left\{ \sqrt{\frac{kd_{\max}d_1 \ldots d_k}{n}}, \ \frac{kd_1 \ldots d_k}{n} \right\} \log^{k+2} d_{\max}.
\]
References.


