On Estimating Rank-One Spiked Tensors in the Presence of Heavy Tailed Errors

Arnab Auddy and Ming Yuan

Abstract— **In this paper, we study the estimation of a rank-one spiked tensor in the presence of heavy tailed noise. Our results highlight some of the fundamental similarities and differences in the tradeoff between statistical and computational efficiencies under heavy tailed and Gaussian noise. In particular, we show that, for pth order tensors, the tradeoff manifests in an identical fashion as the Gaussian case when the noise has finite $4(p-1)$th moment. The difference in signal strength requirements, with or without computational constraints, for us to estimate the singular vectors at the optimal rate, interestingly, narrows for noise with heavier tails and vanishes when the noise only has finite fourth moment. Moreover, if the noise has less than fourth moment, tensor SVD, perhaps the most natural approach, is suboptimal even though it is computationally intractable. Our analysis exploits a close connection between estimating the rank-one spikes and the spectral norm of a random tensor with iid entries. In particular, we show that the order of the spectral norm of a random tensor can be precisely characterized by the moment of its entries, generalizing classical results for random matrices. In addition to the theoretical guarantees, we propose estimation procedures for the heavy tailed regime, which are easy to implement and efficient to run. Numerical experiments are presented to demonstrate their practical merits.**

Index Terms— **Tensor PCA, tensor norm, higher order SVD, robust covariance estimation.**

I. Introduction

Singular value decomposition (SVD) and principal component analysis (PCA) are among the most commonly used procedures in multivariate data analysis. See, e.g., [1], [2]. By seeking low rank approximations to a data matrix, they allow us to reduce the dimensionality of the data, and oftentimes serve as a useful first step to capture the essential features in the data. While both were first developed for the analysis of data matrices, extensions to higher order tensors have also been developed in recent years. See, e.g., [3]–[5]. More generally, low rank tensor methods have exploded in popularity in numerous areas involving high dimensional data analysis. See [6]–[9] for recent reviews.

To fix ideas, consider a rank-one spiked tensor model

$$\mathscr{X} = \lambda \mathbf{u}_1 \circ \mathbf{u}_2 \circ \cdots \circ \mathbf{u}_p + \mathscr{E}, \tag{1}$$

where the "singular value" $\lambda \geq 0$ is a scalar, and "singular vectors" $\mathbf{u}_k$s are unit length vectors in $\mathbb{R}^d_k$, and $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ is a noise tensor whose entries are independent and identically distributed random variables with zero mean and unit variance. To fix ideas, we assume in this section that the tensors are "nearly cubic" in that there exists a constant $C > 0$ such that $d/C \leq d_k \leq Cd$ for all k, although our main results are derived and stated for any general $d_1, \ldots, d_p$. The goal is to estimate the singular vectors after observing $\mathscr{X}$ in a high dimensional setting where d is large. In particular, the special case when the noise tensor $\mathscr{E}$ consists of independent standard normal entries has attracted much attention in recent years, and an intriguing gap in statistical efficiencies with or without computational constraints is observed. It can be shown that tensor SVD that seeks the best rank-one approximation to $\mathscr{X}$ yields a consistent estimate of the singular vectors whenever $\lambda \gg d^{1/2}$. Hereafter, we say an estimate $\widehat{\mathbf{u}}_k$ of $\mathbf{u}_k$ is consistent iff $\sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) \to 0$ as $d \to \infty$ where $\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k)$ is the angle between two vectors $\widehat{\mathbf{u}}_k$ and $\mathbf{u}_k$ taking value in $[0, \pi/2]$. It is worth pointing out that computing the best rank-one approximation is known to be NP hard in general (see, e.g., [10], [11]). On the other hand, consistent yet computationally tractable estimates are only known when $\lambda \gtrsim d^{p/4}$. Hereafter $a \gtrsim b$ means that there is a constant C independent of d such that $a \geq Cb$. More specifically, it can be achieved by power iteration initialized with higher order SVD (HOSVD; see, e.g., [3], [12]). While a rigorous argument remains elusive, it is widely conjectured that $d^{p/4}$ is the tight algorithmic threshold below which no consistent estimates can be computed in polynomial time. It is instructive to consider the case when there are independent Gaussian errors, and the signal strength $\lambda \sim d^{\xi}$. These results can then be summarized by the following diagram. When $\xi > 1/2$, the tensor SVD estimate $\widehat{\mathbf{u}}_k^{\rm SVD}$ is consistent, and indeed can be shown to be minimax rate optimal. Meanwhile, we only know of polynomial time computable estimators that are consistent if $\xi > p/4$. The shaded region between $\xi = 1/2$ and $\xi = p/4$ in Figure 1 therefore signifies the tradeoff between statistical and computational efficiencies.

See, e.g., [5], [13]–[16] among many others. These observations can also be generalized beyond rank-one signals. See, e.g., [17], [18].

The Gaussian, or more generally subgaussian, assumption on the noise tensor $\mathscr{E}$, however, could be too restrictive in practice and neglecting departures from such assumptions could lead to erroneous results. See [19], [20] for detailed discussions in this context. For example, [21] showed how

Manuscript received 20 July 2021; revised 13 April 2022; accepted 6 July 2022. Date of publication 18 July 2022; date of current version 22 November 2022. This work was supported by the NSF under Grant DMS-2015285 and Grant DMS-2052955. *(Corresponding author: Arnab Auddy.)*

The authors are with the Department of Statistics, Columbia University, New York, NY 10027 USA (e-mail: aa4238@columbia.edu).

Communicated by S. Jalali, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3191883.

Digital Object Identifier 10.1109/TIT.2022.3191883

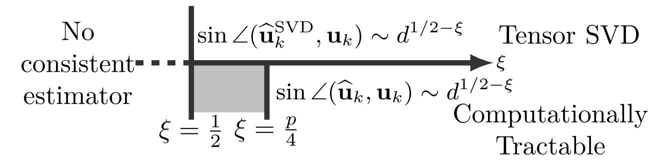
No consistent estimator
$\sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \sim d^{1/2-\xi}$
Tensor SVD
ξ
$\sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) \sim d^{1/2-\xi}$
Computationally Tractable
$\xi = \frac{1}{2}$
$\xi = \frac{p}{4}$

Fig. 1. Tradeoff between statistical and computational efficiencies in estimating spiked rank-one tensors under Gaussian noise (when $d/C \le d_1, \ldots, d_p \le Cd$).

using Gaussian model based methods lead to very high false positive rate in fMRI studies. [22] and [23] observed similar phenomena in genomic studies and anomaly detection respectively. In addition to practical applications, it is also valuable, from the optimization perspective, to consider errors beyond the Gaussian model. The tensor SVD problem has received a lot of attention recently, but most of the existing work focuses on the Gaussian error model. The analogous problem is well studied in random matrix theory. Performance of spectral method for heavy tailed matrices have been studied in the classical works of [24], [25] and more recently in [26] and [27].

Unfortunately, very little is known about the fundamental limit for estimating the rank-one spikes and the effect of computational constraints in the presence of heavy-tailed noise. A notable exception is the recent work of [28] who developed polynomial time algorithms to recover the singular vectors $\mathbf{u}_k$s through self avoiding walks and random coloring. They assume that the singular vectors are randomly sampled and therefore provide an average case analysis of their algorithms. More specifically, for third order tensor ($p = 3$), if the entries of the error tensor has finite second moment, then their algorithm produces weak recovery when $\lambda \gtrsim d^{3/4}$. Moreover, their algorithm yields consistent estimates of the singular vectors if higher order moment conditions, e.g., finite 12th moment, are satisfied. Our work is inspired by this earlier development and aims at developing more practical algorithms for estimating spiked rank-one tensors and precise characterization of how the tradeoff between computational and statistical efficiency manifests beyond subgaussian errors. More specifically, we show that there are polynomial time computable estimates of $\mathbf{u}_k$ that are not only consistent but also rate optimal whenever $\lambda \gg d^{p/4} \cdot \mathrm{polylog}(d)$ where $\mathrm{polylog}(d)$ is a certain polynomial of $\log d$.

Our work is thus among the first ones to highlight the behavior of the tensor SVD optimization problem under independent heavy-tailed errors. For example, existing results on estimation rates might lead one to believe that we require $\lambda \gg \|\mathscr{E}\|$ and the best possible estimation rate is $\|\mathscr{E}\|/\lambda$. We show that this is not the case and it is indeed possible to get the improved rate of $\sqrt{d}/\lambda$, provided that λ is at the correct computational threshold. Further, we control the relevant quantities of the random error tensor that strongly affect the estimation procedure, and provide ways to reduce their influence. This leads to a much more tractable algorithm than the one provided by [28].

The most natural approach to, and a useful benchmark for, estimating $\mathbf{u}_k$s is the tensor SVD. Denote by $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s the tensor SVD estimates of $\mathbf{u}_k$s. Recall that we make the simplifying assumption $d/C \le d_k \le Cd$ for all $1 \le k \le p$. We prove that if the entries of $\mathscr{E}$ have finite αth moment for some $\alpha > 4$, then with high probability,

$$\max_{1\le k\le p} \sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) = O_p\left(\frac{\sqrt{d}}{\lambda}\right), \tag{2}$$

as $d \to \infty$, provided that

$$\lambda \gtrsim \left[d^{1/2}(\log d)^{1/2} + d^{(p-1)/\alpha+1/4}(\log d)^{3/2}\right].$$

The above requirement on the signal-to-noise ratio can also be shown to be optimal, up to the logarithmic factor. More specifically, if the entries of $\mathscr{E}$ do not have finite αth moment, then

$$\sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \to_p 1,$$

for any

$$\lambda \lesssim \left(d^{1/2} + d^{(p-1)/\alpha+1/4}\right).$$

A version of this continues to be true for general dimensions $d_1, \ldots, d_p$. See Section II. It is worth noting that the bounds on λ highlights the intuitive facts that, under the same moment condition, estimating $\mathbf{u}_k$s tends to be harder for higher order tensors, e.g., larger p; and for tensors of the same order, estimating $\mathbf{u}_k$s tends to be easier with higher order moment, e.g., larger α.

It is, however, well known that the tensor SVD is computationally infeasible in general. A common strategy to alleviate the computational expenses of the tensor SVD is through power iteration with spectral initialization. The rationale behind this is the presumptive optimality of the tensor SVD. A good initialization may ensure the resulting estimate, computable in polynomial time, inherits such optimality. We show that this is indeed the case: if $\lambda \gtrsim d^{p/4}$ (more generally $\lambda \gtrsim \max\{\sqrt{d_{\max}}, (d_1 d_2 \ldots d_p)^{1/4}\}$ when the dimensions are unequal), then this yields a polynomial time computable estimate $\widehat{\mathbf{u}}_k$ such that

$$\max_{1\le k\le p} \sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) = O_p\left(\frac{\sqrt{d}}{\lambda}\right).$$

The signal strength requirement for polynomial time computable methods matches that under Gaussian noise and is strictly stronger than that for the tensor SVD estimate. Therefore, the tradeoff between computational and statistical efficiency remains. In particular, if we consider the case when $\lambda \sim d^{\xi}$, then our observations can be summarized by the diagram of Figure 2. The gap between the signal-to-noise ratio requirement for tensor SVD and polynomial computable estimators is the same as in the Gaussian case when $\alpha \ge 4(p-1)$ but narrows as α decreases to 4.

A more intriguing phenomenon occurs when the entries of $\mathscr{E}$ only has finite αth moment for some $2 < \alpha < 4$. In this situation, we prove that (2) holds if

$$\lambda \gtrsim d^{\frac{p-1}{\alpha}+\frac{1}{2}}(\log d)^{3/2}$$

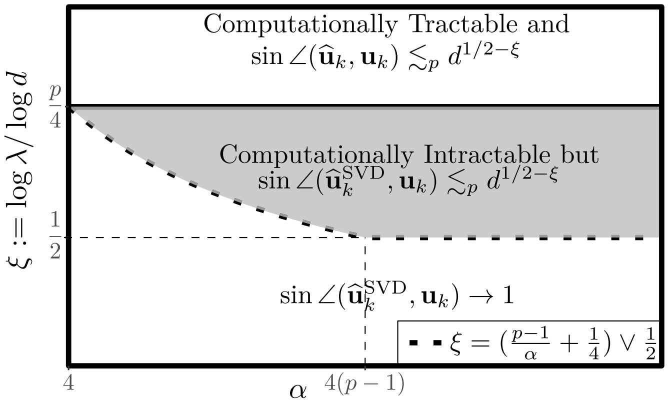

Fig. 2. Tradeoff between statistical and computational efficiencies in estimating spiked rank-one tensors when the noise has more than fourth moments (when $d/C \leq d_1, \ldots, d_p \leq Cd$).

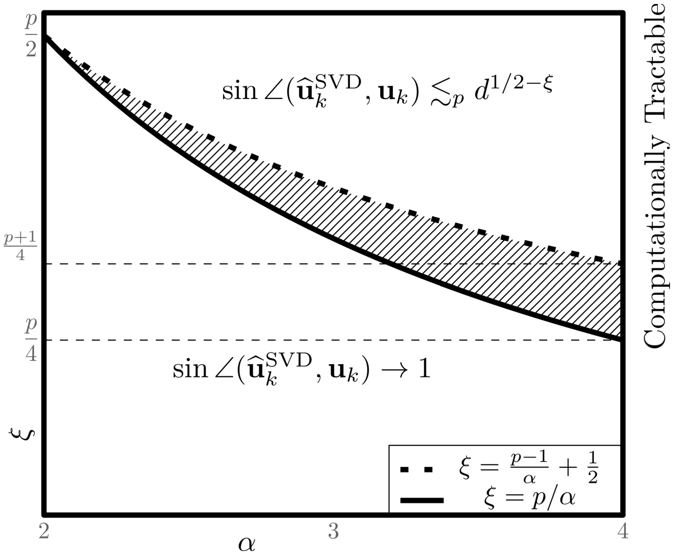

Fig. 3. Tradeoff between statistical and computational efficiencies in estimating spiked rank-one tensors when the noise does not have fourth moments (when $d/C \leq d_1, \ldots, d_p \leq Cd$).

and the tensor SVD estimate $\widehat{\mathbf{u}}_k^{\rm SVD}$ is asymptotically perpendicular to $\mathbf{u}_k$ if

$$\lambda \lesssim d^{p/\alpha}.$$

This can be summarized by the diagram of Figure 3. We state a more general result in Section II. Interestingly, the tensor SVD is actually suboptimal in this case and there is an alternative estimator that is both computationally tractable and can attain the optimal rate of convergence whenever

$$\lambda \gtrsim d^{p/4}(\log d)^{1/4}.$$

More generally, we require the singular value to be $\lambda \gtrsim \max\{\sqrt{d_{\max}}, (d_1 \ldots d_p)^{1/4} \log(d_{\max})^{1/4}\}$ when the dimensions are possibly of different magnitude. Due to the suboptimality of tensor SVD, it is doubtful if power iteration would work when $\alpha < 4$. To this end, we consider a different estimating strategy. More specifically, our techniques are based on recent developments in the theory of robust estimation of the mean in the presence of heavy tailed errors. These works derive estimators with subgaussian concentration, inspired from the pioneering work of [29]. The key idea is to reduce the adverse effect of heavy tails through an influence function, and can be extended to matrix estimation. For covariance matrix estimation, [30] and [31] were some of the first works in this area, although both these approaches involved optimizing over a d-dimensional ϵ-net and thus having exponential time complexity. [32] have similar results with polynomial time, but they too require an extensive search for tuning parameters. We will instead use results on spectrum truncated estimators applied to covariance estimation. [33] described one such method for robust PCA through smooth truncation, based on which [34] and [35] provided more tractable procedures and general results.

Our results are obtained by exploiting close connections between estimating the rank-one spikes and the spectral norm of a random tensor of iid entries. We show that the order of the spectral norm of a random tensor can be precisely characterized by the moment of its entries, which might be of independent interest. In particular, our result indicates that, up to a logarithmic factor, the norm of the random tensor $\|\mathscr{E}\|$ is of the order $\sqrt{d}$ if and only if its entries have finite $4(p-1)$th moment. This can be viewed as a generalization of the classical results for random matrices ([24], [36]). In deriving these bounds, we used techniques developed for random matrices by [37] and improved moment bounds of random tensors established earlier by [38].

The rest of the paper is organized as follows. We first develop probabilistic bounds for the spectral norm of a random tensor of iid entries and use these tools to study the performance of the tensor SVD in Section II. Polynomial time computable estimation schemes are given in Sections III and IV for $\alpha \geq 4$ and for $\alpha \geq 2$ respectively. To corroborate our theoretical development, Section V provides simulation studies to further demonstrate the practical merits of the proposed methods. We conclude with a few remarks on the implications and future directions in Section VI. All proofs are given in Section VII.

A. Notation

Alongside the standard notation for vectors, matrices and tensors, we will use the following special notation. We will use $\otimes$ to denote the Kronecker product, and $\circ$ to denote the outer product. For $d \in \mathbb{N}$, we write $[d]$ to mean the set $\{1, \ldots, d\}$. For dimensions $d_1, \ldots, d_p \in \mathbb{N}$, we will denote $d_{-k} = \prod_{q \neq k} d_q$. Finally we will use the maximum and the geometric mean

$$d_{\max} = \max\{d_1, \ldots, d_p\}, \quad \bar{d}_G = \left(\prod_{k=1}^{p} d_k\right)^{1/p}. \qquad (3)$$

II. Tensor SVD and Spectral Norm of Random Tensors

The most natural approach to estimating the singular vectors is via the tensor SVD. In particular, let

$$(\widehat{\mathbf{u}}_k^{\rm SVD} : 1 \leq k \leq p) = \operatorname{argmax}_{\mathbf{a}_k \in \mathbb{S}^{d_k - 1}} \langle \mathscr{X}, \mathbf{a}_1 \circ \cdots \circ \mathbf{a}_p \rangle, \qquad (4)$$

Here $\mathbb{S}^{d_k-1}$ is the unit sphere in $\mathbb{R}^{d_k}$. It is well known that the tensor SVD can be equivalently characterized the best rank-one apprpoximation to $\mathscr{X}$ in that

$$\begin{aligned}&(\widehat{\lambda}^{\mathrm{SVD}}, \widehat{\mathbf{u}}_k^{\mathrm{SVD}} : 1 \le k \le p) \\ &= \operatorname{argmax}_{\gamma\in\mathbb{R}, \mathbf{a}_k \in \mathbb{S}^{d_k-1}} \|\mathscr{X} - \gamma \mathbf{a}_1 \circ \cdots \circ \mathbf{a}_p\|_{\mathrm{HS}},\end{aligned}$$

where $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert-Schmidt or Frobenius norm. See, e.g., [39]. The performance of these singular vector estimates is closely related to the spectral norm of the noise tensor:

$$\|\mathscr{E}\| := \max_{\mathbf{a}_k \in \mathbb{S}^{d_k-1}} \langle \mathscr{E}, \mathbf{a}_1 \circ \cdots \circ \mathbf{a}_p \rangle.$$

For example, it is known that

$$\max_{1\le k\le p} \sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \lesssim \frac{\|\mathscr{E}\|}{\lambda}, \tag{5}$$

so that $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s are consistent whenever $\lambda \gg \|\mathscr{E}\|$. See, e.g., [18]. To this end, we shall first study the spectral norm of a random tensor consisting of independent and identically distributed entries.

A. Norm of Random Tensors

The problem of bounding the spectral norm of a random tensor is well-studied in the matrix case, i.e., $p = 2$. In particular, [36] showed that if $\mathscr{E}$ is an iid ensemble, then $\|\mathscr{E}\|$ is of the order $\sqrt{d_{\max}}$ if and only if its entries have finite (weak) fourth moment. In other words, when $p = 2$ and the entries of $\mathscr{E}$ have finite fourth moment, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s are consistent if and only if $\lambda \gg \sqrt{d_{\max}}$. A couple of questions arise naturally. First, is there similar characterization of $\|\mathscr{E}\|$ for higher order tensors? And what happens if the entries of $\mathscr{E}$ have only αth moment for $2 \le \alpha < 4$? The next result aims to settle the first question.

Theorem 2.1: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E with mean zero and variance σ^2. Then there exists a constant $C_p > 0$ depending on p only such that for any $\alpha \ge 4$, $\mathbb{E}|E|^\alpha < \infty$ implies that, with probability at least $1 - d_{\max}^{-\alpha/4+1}$,

$$\|\mathscr{E}\| \le C_p\sigma \left[d_{\max}^{\frac{1}{2}} + \left(\bar{d}_G\right)^{\frac{p}{\alpha}} (d_{\max})^{\frac{1}{4}-\frac{1}{\alpha}} (\log d_{\max})^{3/2} \right].$$

Conversely, there exists another constant $C_p' > 0$ depending on p only such that $\mathbb{E}|E|^\alpha = \infty$ implies that

$$\|\mathscr{E}\| \ge C_p'\sigma \left[d_{\max}^{\frac{1}{2}} + \left(\bar{d}_G\right)^{\frac{p}{\alpha}} (d_{\max})^{\frac{1}{4}-\frac{1}{\alpha}} \right], \text{ almost surely}$$

as $d_{\max} \to \infty$. Here $d_{\max} = \max\{d_1, \ldots, d_p\}$, and $\bar{d}_G = \left(\prod_{k=1}^p d_k\right)^{1/p}$.

The lower and upper bounds of Theorem 2.1 match up to the logarithmic factor when $4 \le \alpha \le 4(p-1)$ and match up to constants whenever $\alpha > 4(p-1)$. In particular, $\|\mathscr{E}\|$ is of the order $\sqrt{d_{\max}}$, if and only if its entries have finite $4(p-1)$th moment. This can be viewed as a generalization of the classical result for $p = 2$ from [36]. For higher order tensors ($p > 2$), the precise order of $\|\mathscr{E}\|$ depends on the value of α for $4 \le \alpha < 4(p-1)$. Consider, for example, $p = 3$. Then $\|\mathscr{E}\|$ is of the same order as that of an iid Gaussian ensemble, up to at most a logarithmic factor, as soon as E has finite eighth moment. Yet, if E only has finite αth moment for $4 \le \alpha < 8$, then $\|\mathscr{E}\|$ depends on the exact value of α, and decreases as α increases.

The next result complements Theorem 2.1 and deals with the case when $2 \le \alpha < 4$.

Theorem 2.2: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E with mean zero and variance σ^2. There exist constants $C_p, C_p' > 0$ depending on p only such that for any $2 \le \alpha < 4$, $\mathbb{E}|E|^\alpha < \infty$ implies that, with probability at least $1 - d_{\max}^{-\alpha/2+1}$,

$$\|\mathscr{E}\| \le C_p\sigma \left(\bar{d}_G\right)^{\frac{p}{\alpha}} (d_{\max})^{\frac{1}{2}-\frac{1}{\alpha}} (\log d_{\max})^{3/2}.$$

Conversely, if $\mathbb{E}|E|^\alpha = \infty$ then

$$\|\mathscr{E}\| \ge C_p'\sigma\sqrt{d_{\max}} + C_p'\sigma(\bar{d}_G)^{\frac{p}{\alpha}}, \quad \text{almost surely}$$

as $d_{\max} \to \infty$. Here $d_{\max} = \max\{d_1, \ldots, d_p\}$ and $\bar{d}_G = \left(\prod_{k=1}^p d_k\right)^{1/p}$.

Note that there is a gap between the upper bound and lower bound in Theorem 2.2 beyond the logarithmic factor. While it is plausible that this is the result of our proof technique, it remains a possibility that this may point to something more fundamental.

We want to mention here that for $p = 2$, i.e., matrices, these bounds can be compared to the matrix norm bounds from [25]. Our bounds are worse by the $(\log d)^{3/2}$ factor.

B. Convergence Rates for Tensor SVD

In light of (5), Theorems 2.1 and 2.2 immediately imply the consistency of $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s when

$$\begin{aligned}&\lambda \gg \lambda_{\mathrm{crit}}(\mathbf{d};\alpha) \\ &:= \begin{cases} d_{\max}^{\frac{1}{2}} & \text{if } \alpha > \tilde{p} \\ d_{\max}^{\frac{1}{2}} + \left(\bar{d}_G\right)^{\frac{p}{\alpha}} (d_{\max})^{\frac{1}{4}-\frac{1}{\alpha}} (\log d_{\max})^{\frac{3}{2}} & \text{if } 4 \le \alpha \le \tilde{p} \\ \left(\bar{d}_G\right)^{\frac{p}{\alpha}} (d_{\max})^{\frac{1}{2}-\frac{1}{\alpha}} (\log d_{\max})^{\frac{3}{2}} & \text{if } 2 \le \alpha < 4 \end{cases}\end{aligned} \tag{6}$$

where $\tilde{p} = 4(p-1)$.

Recall that $d_{\max} = \max\{d_1, \ldots, d_p\}$ and $\bar{d}_G = \left(\prod_{k=1}^p d_k\right)^{1/p}$. In fact, under this condition of the signal-to-noise ratio, much stronger statement can be made and in fact, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s can be shown to be rate optimal:

Theorem 2.3: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E with mean zero, variance one, and finite αth moment, e.g., $\mathbb{E}|E|^\alpha < \infty$ for some $\alpha \ge 2$. Then there exist a numerical constant $C > 0$ and another constant C_p depending on p only such that if $\lambda \ge C\lambda_{\mathrm{crit}}(\mathbf{d};\alpha)$, then

$$\max_{1\le k\le p} \sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \le C_p \frac{\sqrt{d_{\max}}}{\lambda},$$

with probability tending to one as $d_{\max}$ increases.

For comparison, under Gaussian noise, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$ converges to $\mathbf{u}_k$ at the optimal rate of $\sqrt{d_{\max}}/\lambda$ as soon as $\lambda > C\sqrt{d_{\max}}$ for some constant $C > 0$. Theorem 2.3 shows that the same is true, up to some constant factor, when the entries of $\mathscr{E}$ has finite $4(p-1)$th moment. However, when $\alpha < 4(p-1)$, the rate $\sqrt{d_{\max}}/\lambda$ can only be achieved when λ is much larger than that required with Gaussian errors. Nonetheless the following result shows that when $\alpha > 4$ these requirements are indeed optimal, up to a logarithmic factor, and therefore highlight a fundamental difference in behavior of tensor SVD with heavy tailed and Gaussian noise.

Theorem 2.4: Assume, without loss of generality, that $d_1 \geq \cdots \geq d_p$ and define $d_{\max}$ and $\bar{d}_G$ as in (3). Let $\mathscr{E} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E such that $\mathbb{E}|E|^\alpha = \infty$ for some $4 < \alpha < 4(p-1)$ yet $\mathbb{E}|E|^\beta < \infty$ for some β such that $(d_2 d_3 \ldots d_{p-1})^{1/\beta} \leq (d_2 d_3 \ldots d_p)^{1/\alpha}$ and $\beta > 4$. If $\lambda < C\left(\bar{d}_G\right)^{\frac{p}{\alpha}}(d_{\max})^{\frac{1}{4}-\frac{1}{\alpha}}$ for any constant $C > 0$, then for any constant $0 < C_0 < 1$,

$$\min_{1 \leq k \leq p} \sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \geq C_0,$$

with probability tending to one, as $d_1 \to \infty$. Similarly, suppose that $\mathbb{E}|E|^\alpha = \infty$ for some $2 < \alpha < 4$ and $\mathbb{E}|E|^\beta < \infty$ for some β such that $(d_2 d_3 \ldots d_{p-1})^{1/\beta} \leq (d_2 d_3 \ldots d_p)^{1/\alpha}$ and $\beta > 2$. If $\lambda < C\sqrt{d_{\max}} + \left(\bar{d}_G\right)^{p/\alpha}$ for any constant $C > 0$, then for any constant $0 < C_0 < 1$,

$$\min_{1 \leq k \leq p} \sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \mathbf{u}_k) \geq C_0$$

with probability tending to one, as $d_1 \to \infty$.

For concreteness, consider a continuous distribution symmetric about 0 whose survival function is given by

$$\bar{F}(x) := 1 - F(x) = x^{-\alpha}L(x), \qquad x > 0$$

where $L(x)$ is slowly varying function at $+\infty$ in that

$$L(x) > 0 \quad \text{and} \quad \lim_{x\to\infty}\frac{L(tx)}{L(x)} = 1, \qquad t > 0.$$

For such distributions, α is often referred to as their tail index. It is clear that for $E \sim F$, $\mathbb{E}(|E|^q) = \infty$ if and only if $q \geq \alpha$. In light of Theorem 2.4, when $\alpha > 4$, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$ is inconsistent if $\lambda \lesssim \max\left\{(\bar{d}_G)^{p/\alpha}d_{\max}^{1/4-1/\alpha}, \sqrt{d_{\max}}\right\}$; on the other hand, when $2 < \alpha < 4$, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$ is inconsistent if $\lambda \lesssim \max\left\{\sqrt{d_{\max}}, (\bar{d}_G)^{p/\alpha}\right\}$. Conversely as a result of Theorem 2.3, $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$ converges to $\mathbf{u}_k$ at the optimal rate if $\lambda \gtrsim \lambda_{\mathrm{crit}}(\mathbf{d}; \alpha - \epsilon)$ for any $\epsilon > 0$.

Interestingly, perhaps also surprisingly at the first sight, the inferior signal strength requirement for estimating the singular vectors under heavy-tailed noise is only a limitation of the tensor SVD and not a fundamental barrier in general. We now show that it is possible to improve the tensor SVD via a different estimation strategy at least when the signal-to-noise ratio is sufficiently high.

III. Power Iteration With Spectral Initiation

One of the chief challenges with the tensor SVD is the computational cost. It is well known that computing the best rank-one approximation (4) is NP hard (e.g. [10], [11]) so that it is infeasible to compute $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s for large d_k. A common strategy to overcome this difficulty is to apply power iteration with spectral initialization, which has been shown to yield an estimator that is both polynomial time computable and rate optimal in the presence of Gaussian error. See, e.g., [5], [13]. We shall now show that this strategy continues to work whenever $\alpha > 4$.

Recall that the first order condition yields that $\widehat{\mathbf{u}}_k^{\mathrm{SVD}}$s satisfies

$$\mathscr{X} \times_{j\neq k} \widehat{\mathbf{u}}_j^{\mathrm{SVD}} \propto \widehat{\mathbf{u}}_k^{\mathrm{SVD}}, \qquad 1 \leq k \leq p.$$

Motivated by this property, we shall consider estimating $\mathbf{u}_k$ through power iteration:

$$\mathbf{x}_k^{[t+1]} = \frac{\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]}}{\left\|\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]}\right\|}, \tag{7}$$

with initial estimates $\mathbf{x}_j^{[0]}$s. For this to work, we first need to be able to find a "reasonably good" initial estimate $\mathbf{x}_j^{[0]}$ that can be efficiently computed. This is usually done through HOSVD.

More specifically, denote by $\mathsf{Mat}_k : \mathbb{R}^{d_1\times\cdots\times d_p} \to \mathbb{R}^{d_k \times d_{-k}}$ the operator that collapses all indices other than the kth one of a pth order tensor and therefore converts it into a $d_k \times d_{-k}$ matrix. Here we use the notation $d_{-k} = \prod_{q\neq k} d_q$. Write

$$\mathscr{T} = \lambda\mathbf{u}_1 \circ \cdots \circ \mathbf{u}_p.$$

It is not hard to see that

$$\mathsf{Mat}_k(\mathscr{T}) = \lambda\mathbf{u}_k(\mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_{k-1} \otimes \mathbf{u}_{k+1} \otimes \cdots \otimes \mathbf{u}_p)^\top,$$

where $\otimes$ stands for the Kronecker product so that we can estimate $\mathbf{u}_k$ by the leading left singular vector, denoted by $\widehat{\mathbf{u}}_k^{\mathrm{Mat}}$, of $\mathsf{Mat}_k(\mathscr{X})$. Observe that

$$\mathbb{E}\left[\mathsf{Mat}_k(\mathscr{X})\mathsf{Mat}_k(\mathscr{X})^\top\right] = \lambda^2\mathbf{u}_k \circ \mathbf{u}_k + d_{-k}\mathbb{I}_{d_k},$$

and $\widehat{\mathbf{u}}_k^{\mathrm{Mat}}$ is the leading eigenvector of the matrix $\mathsf{Mat}_k(\mathscr{X})\mathsf{Mat}_k(\mathscr{X})^\top - d_{-k}\mathbb{I}_{d_k}$. By Davis-Kahan Theorem, we have

$$\begin{aligned}
&\sin\angle(\mathbf{u}_k, \widehat{\mathbf{u}}_k^{\mathrm{Mat}})\\
\leq&\frac{2}{\lambda^2}\big\|\mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{E})^\top - d_{-k}\mathbb{I}_{d_k} +\\
&+ \mathsf{Mat}_k(\mathscr{T})\mathsf{Mat}_k(\mathscr{E})^\top + \mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{T})^\top\big\|\\
\leq&\ \frac{2\|\mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{E})^\top - d_{-k}\mathbb{I}_{d_k}\|}{\lambda^2}\\
&+ \frac{4\|\mathsf{Mat}_k(\mathscr{T})\mathsf{Mat}_k(\mathscr{E})^\top\|}{\lambda^2}.
\end{aligned}$$

Following Bai-Yin's law, we then have

Proposition 3.1: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E with mean zero, variance one and $\mathbb{E}|E|^\alpha < \infty$ for some $\alpha \geq 4$. Then

$$\sin\angle(\widehat{\mathbf{u}}_k^{\mathrm{Mat}}, \mathbf{u}_k) = O_p\left(\frac{d_k + \left(\bar{d}_G\right)^{p/2} + \lambda\sqrt{d_k}}{\lambda^2}\right),$$

as $d_{\max} \to \infty$, for $1 \le k \le p$. Here $d_{\max}$ and $\bar{d}_G$ are as defined in (3).

Proposition 3.1 indicates that $\widehat{\mathbf{u}}_k^{\text{Mat}}$s are consistent as soon as $\lambda \gg d_{\max}^{1/2} + (\bar{d}_G)^{p/4}$. It is worth comparing this requirement with that of $\widehat{\mathbf{u}}_k^{\text{SVD}}$s: $\lambda \gg d_{\max}^{1/2} + d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{p/\alpha}$. See Theorem 2.4. The former is more restrictive since $\alpha \ge 4$. As in the Gaussian noise case, this gap is likely a display of the tradeoff between computational and statistical efficiencies: $\widehat{\mathbf{u}}_k^{\text{Mat}}$ is computationally tractable yet $\widehat{\mathbf{u}}_k^{\text{SVD}}$ in general is not. On the other hand, the convergence rate for $\widehat{\mathbf{u}}_k^{\text{Mat}}$ is inferior to that of $\widehat{\mathbf{u}}_k^{\text{SVD}}$. However, we can improve upon $\widehat{\mathbf{u}}_k^{\text{Mat}}$s by using $\widehat{\mathbf{u}}_j^{\text{Mat}}$s in place of $\mathbf{x}_j^{[0]}$s in (7) to get an updated estimate.

To see how this works, write

$$\mathbf{x}_j^{[t]} = \sqrt{1-\rho_j^2}\,\mathbf{u}_j + \rho_j \mathbf{v}_j$$

where $\mathbf{v}_j$ is a unit length vector perpendicular to $\mathbf{u}_j$. Then

$$\begin{aligned}
&\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]} \\
=&\lambda \left(\prod_{j\neq k} \sqrt{1-\rho_j^2}\right) \mathbf{u}_k \\
&+ \sum_{A\subset([p]\setminus\{k\})} \left(\prod_{j\in A}\sqrt{1-\rho_j^2} \prod_{j\notin A\cup\{k\}} \rho_j\right) \cdot \\
&\qquad \mathscr{E} \times_{j\in A} \mathbf{u}_j \times_{j\notin A\cup\{k\}} \mathbf{v}_j .
\end{aligned}$$

Note that the second term on the righthand side can be bounded by, up to a constant, $\|\mathscr{E}\|$. In light of Proposition 3.1, this implies that, if ρ_j are uniformly bounded away from 1, then

$$\sin\angle\left(\mathbf{x}_k^{[t+1]}, \mathbf{u}_k\right) = O_p\left(\frac{\|\mathscr{E}\|}{\lambda}\right).$$

In particular, in the case of Gaussian errors, $\|\mathscr{E}\| = O_p(\sqrt{d_{\max}})$ so that we can conclude that

$$\sin\angle\left(\mathbf{x}_k^{[1]}, \mathbf{u}_k\right) = O_p\left(\frac{\|\mathscr{E}\|}{\lambda}\right)$$

suggesting that a single iteration with $\mathbf{x}_k^{[0]} = \widehat{\mathbf{u}}_k^{\text{Mat}}$ ($k = 1, \dots, p$) leads to rate optimal estimates of $\mathbf{u}_k$. The same technique can be applied whenever $\alpha > 4(p-1)$ thanks to Theorem 2.1. The argument, however, breaks down when $\alpha < 4(p-1)$ and a single iteration no longer suffices. Nonetheless, a more careful analysis shows that the performance keeps improving with more iterations and $O(\log d_{\max})$ number of iterations can yield a rate optimal of $\mathbf{u}_k$s.

Proposition 3.2: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent copies of a random variable E with mean zero, variance one and $\mathbb{E}|E|^\alpha < \infty$ for some $\alpha > 4$. There exist constants $C_1, C_2 > 0$ such that if $\lambda > C_1 d_{\max}^{1/2} + C_1\left(\bar{d}_G\right)^{p/4}$ and $\rho^{[t]} < 1$, then

$$\rho_k^{[t+1]} \le C_2(\rho_k^{[t]})^2 \frac{\|\mathscr{E}\|}{\lambda} + O_p\left(\frac{\sqrt{d_k}}{\lambda}\right),$$

where

$$\rho_k^{[t]} = \sin\angle(\mathbf{x}_k^{[t]}, \mathbf{u}_k).$$

In light of Propositions 3.1 and 3.2, we can estimate $\mathbf{u}_k$ by running power iterations (7) with initialization

$$\mathbf{x}_k^{[0]} = \widehat{\mathbf{u}}_k^{\text{Mat}}, \qquad k = 1, \dots, p.$$

And

$$\sin\angle(\mathbf{x}_k^{[T]}, \mathbf{u}_k) = O_p\left(\frac{\sqrt{d_k}}{\lambda}\right), \qquad \text{as } d \to \infty,$$

for $1 \le k \le p$ and $T \gtrsim \log\left(\bar{d}_G\right)$ provided that $\lambda \ge C d_{\max}^{1/2} + C\left(\bar{d}_G\right)^{p/4}$ for a sufficiently large constant $C > 0$. This proves that

Theorem 3.3: Assume that the entries of $\mathscr{E}$ are independent and identically distributed with zero mean, unit variance and finite αth moment for some $\alpha > 4$. There exist constants $C_1, C_2 > 0$ such that if $\lambda > C_1 d_{\max}^{1/2} + C_1\left(\bar{d}_G\right)^{p/4}$, then there is a polynomial time computable estimator $\widehat{\mathbf{u}}_k$ ($k = 1, \dots, p$) obeying

$$\sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) \le \frac{C_2\sqrt{d_k}}{\lambda},$$

with probability at least $1 - d_k^{-1}$ for $1 \le k \le p$. Here $d_{\max}$ and $\bar{d}_G$ are as defined in (3).

For this strategy to work we need $\lambda > \|\mathscr{E}\|$. However, in light of Theorem 2.2, this would require a higher signal-to-noise ratio than $\left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ when $\alpha < 4$. It turns out that while the vanilla power iteration may not work for smaller αs, it is possible to attain both statistical and computational efficiencies as long as $\lambda \gtrsim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ for any $\alpha \ge 2$.

IV. Tractable Estimation for All $\alpha \ge 2$

As indicated in Theorem 3.3, HOSVD and power iteration yields a consistent estimator under the signal strength requirement $\lambda \gtrsim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ only if the entries of $\mathscr{E}$ have finite fourth moment. This can no longer be successful when $\lambda \sim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ and $\alpha < 4$, even without computational considerations, as shown by Theorem 2.4. To resolve this issue, we need to modify both the initialization and the power iteration steps. We first describe a new way for initialization.

A. Initialization by Robust HOSVD

The rationale behind the spectral initialization is that $\mathsf{Mat}_k(\mathscr{X})\mathsf{Mat}_k(\mathscr{X})^\top$ is an unbiased estimate of $\lambda^2\mathbf{u}_k \circ \mathbf{u}_k$. However, this incurs bounding $\|\mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{E})^\top - d_{-k}I\|$ which requires finite fourth moment of the entries of $\mathscr{E}$. To relax this condition, we shall now proceed to estimate $\lambda^2\mathbf{u}_k \circ \mathbf{u}_k$ via a more robust approach that works as long as the entries of $\mathscr{E}$ have finite variance.

In particular, we shall adopt a method first developed by [29] for estimating univariate mean, and later extended by [34] for estimating matrices. It is based on an M-estimation framework where we estimate the common mean $\mathbf{M}$ from some independent, but not necessarily identically distributed, samples $\mathbf{S}_i, i = 1, \dots, n$ by

$$\widehat{\mathbf{S}} = \operatorname{argmin}_{\mathbf{M}} \left[\operatorname{tr}\sum_{j=1}^n \Psi(\theta(\mathbf{S}_j - \mathbf{M}))\right],$$

and θ is a tuning parameter to be specified later. Here, for a function $f : \mathbb{R} \to \mathbb{R}$ and symmetric matrix $\mathbf{M}$ with spectral decomposition $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$,

$$f(\mathbf{M}) = \mathbf{U}\,\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_d))\mathbf{U}^\top.$$

In particular, we shall take a Ψ so that its first derivative $\psi = \Psi'$ is operator Lipschitz and obeys

$$-\log(1 - x + x^2/2) \le \psi(x) \le \log(1 + x + x^2/2).$$

See [29] and [34] for further discussions and various examples.

Recall that

$$\mathsf{Mat}_k(\mathscr{X})\mathsf{Mat}_k(\mathscr{X})^\top = \sum_{i_{-k} \in [d]_{-k}} \mathbf{X}_{i_{-k}}\mathbf{X}_{i_{-k}}^\top$$

where $i_{-k} = (i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_p)$, $[d_{-k}] = [d_1] \times \ldots [d_{k-1}] \times [d_{k+1}] \ldots [d_p]$ and $\mathbf{X}_{i_{-k}}$ is the kth mode fiber of $\mathscr{X}$ with all indices except for the kth one fixed. Note that

$$\mathbb{E}(\mathbf{X}_{i_{-k}}\mathbf{X}_{i_{-k}}^\top) = \lambda^2 w_{i_{-k}}^2 \mathbf{u}_k \circ \mathbf{u}_k + I,$$

where

$$w_{i_{-k}} = \prod_{l \ne k} u_{l i_l}.$$

It is tempting to apply the aforementioned strategy directly to $\{\mathbf{X}_{i_{-k}}\mathbf{X}_{i_{-k}}^\top : i_{-k} \in [d_{-k}]\}$ to estimate $\lambda^2 \mathbf{u}_k \circ \mathbf{u}_k$. There are, however, a couple of obstacles in doing so. Firstly, bounding the variation of $\widehat{\mathbf{S}}$ incurs the second moment of $\mathbf{S}_j$s which can be translated into a requirement on the fourth moment of $\mathscr{X}$. This is exactly what we try to avoid. To this end, we shall instead consider estimating

$$\mathbf{V}_k = \lambda^2 \left[\mathbf{u}_k \circ \mathbf{u}_k - \mathrm{diag}(\mathbf{u}_k \circ \mathbf{u}_k)\right].$$

Note that

$$\left\|\mathbf{V}_k - \lambda^2 \mathbf{u}_k \circ \mathbf{u}_k\right\| = \lambda^2 \|\mathbf{u}_k\|_{\ell_\infty}^2.$$

By Davis-Kahan Theorem, we know that

$$\sin\angle(\mathbf{v}, \mathbf{u}_k) \le \frac{2\left\|\mathbf{V}_k - \lambda^2 \mathbf{u}_k \circ \mathbf{u}_k\right\|}{\|\mathbf{V}_k\|} \le 4\|\mathbf{u}_k\|_{\ell_\infty}^2.$$

Therefore, by assuming that $4\|\mathbf{u}_k\|_{\ell_\infty}^2 < \eta$, a "good" estimate of the leading eigenvector of $\mathbf{V}$ may yield an initial value satisfying the requirement of Proposition 4.2.

Another difficulty is that

$$\mathbf{Y}_{i_{-k}} = \mathbf{X}_{i_{-k}}\mathbf{X}_{i_{-k}}^\top - \mathrm{diag}(\mathbf{X}_{i_{-k}}\mathbf{X}_{i_{-k}}^\top)$$

have different means. To this end, we randomly partition $[d_{-k}]$ into n groups, denoted by $I_1, \ldots, I_n$. This sampling is done through $d_2 \ldots d_p$ samples of $\mathrm{Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$. Let

$$\widehat{\mathbf{V}}_k = \frac{1}{n\theta}\sum_{j=1}^n \psi(\theta \mathbf{S}_j),$$

where

$$\mathbf{S}_j = \sum_{i_{-k} \in I_j} \mathbf{Y}_{i_{-k}}. \tag{8}$$

$\widehat{\mathbf{V}}_k$ can be viewed as a one-step gradient descent for computing $\widehat{\mathbf{S}}$ with initial value 0.

Denote by

$$\mu_1 = \max_{i_{-k} \in [d]^{p-1}} |w_{i_{-k}}|, \qquad \text{and} \qquad \mu_2 = \max_{1 \le k \le p} \|\mathbf{u}_k\|_{\ell_\infty}.$$

And write $\widehat{\mathbf{v}}_k$ the leading eigenvector of $\widehat{\mathbf{V}}_k$. Then we have the following theorem.

Theorem 4.1: Assume that the parameters satisfy $\lambda > C\left(d_{\max}^{\frac{1}{2}} + \left(\bar{d}_G\right)^{\frac{p}{4}}\right)(\log(\bar{d}_G))^{\frac{1}{4}}$ and $\mu_1 \le C^{-1}(\log d_k)^{-1}$ for a sufficiently large constant $C > 0$. If

$$\theta = \sqrt{\frac{8\log(\bar{d}_G)}{\lambda^4/n + d_{\max}^2 + (\bar{d}_G)^p}},$$

then

$$\sin\angle(\widehat{\mathbf{v}}_k, \mathbf{u}_k) \le 2\mu_2^2 + \sqrt{\frac{32\log d_k}{n}} + \frac{(\lambda\sqrt{d_k} + (\bar{d}_G)^{p/2})\sqrt{8\log \bar{d}_G}}{\lambda^2}$$

for $1 \le k \le p$ with probability at least $1 - Cd_k^{-1}\log(\bar{d}_G) - n\exp(-1/Cn\mu_1^2)$. Here $d_{\max}$ and $\bar{d}_G$ are as defined in (3).

The algorithm above effectively does a truncation around $\mathbf{0}$. It is natural that this causes significant bias and leads to a larger deviation term. With more gradient iterations, $\widehat{\mathbf{V}}_k^{(t)}$ becomes an increasingly better approximation to $\mathbf{V}_k$ and reduces the second term of the deviation exponentially fast. We omit details since we intend to use this only for initialization and the performance guarantee given by Theorem 4.1 is sufficient for our purpose.

The theoretical choice of the truncation parameter θ as given above, requires some knowledge of λ. If we instead have some preliminary bounds on λ, we define θ_j as follows by the so-called Lepski method. Let $\mathcal{L} = \{l \in \mathbb{N} : \lambda_{\min} \le \lambda_l = 2^l\lambda_{\min} \le 2\lambda_{\max}\}$. For each λ_l the corresponding truncated estimators $\widehat{\mathbf{V}}_{(l)}$ are defined as above. Then,

$$\theta = \sqrt{\frac{8\log \bar{d}_G}{\lambda_l^4/n + d_{\max}^2 + \left(\bar{d}_G\right)^p}}$$

and

$$l^* = \min\left\{ l \in \mathcal{L} : \forall k \in \mathcal{L},\, k > l, \right.$$
$$\left. \|\widehat{\mathbf{V}}_{(l)} - \widehat{\mathbf{V}}_{(k)}\| \le \frac{\lambda_k^4/n + (\bar{d}_G)^p + d_{\max}^2}{12n} \right\}.$$

Using results from [34], it can be shown that this scheme provides estimates that differ from Theorem 4.1 only by a constant. Notice that in our case we can get a crude upper bound for λ using the Frobenius norm of one of the tensor pieces. On the other hand, we can take $\lambda_{\min} = d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}$, which has to be smaller than the actual λ in order to make a successful recovery. We implicitly assume $\lambda \le \mathrm{poly}(\bar{d}_G)$. Otherwise, if λ is exponentially large, we can detect that through the Frobenius norm of the tensor and the problem is simplified. Writing the tensor as a $d_1 \ldots d_p$ dimensional vector and estimating the mean is enough for this case. Thus we need to take at most $Cp\log(d_{\max})$ values of l. Finally, our simulation results show that a fixed upper bound for λ often suffices and we do not need to estimate it.

B. One Step Power Iteration With Sample Splitting

In light of Theorem 4.1, if $\lambda \sim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ $(\log \bar{d}_G)^{1/4}$, then we can ensure that $\sin\angle(\widehat{\mathbf{v}}_k, \mathbf{u}_k) \le \eta$ for some constant $\eta < 1$ by we take $n = C\log d$. We shall now consider using them in the power iteration. As suggested by Proposition 3.2, for the accuracy to improve from iteration to iteration, it is important that we have $\lambda \gtrsim \|\mathscr{E}\|$. In light of Theorem 2.2, the requirement that $\lambda \sim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)$ cannot ensure that is the case when $\alpha < 4$. It turns out that this requirement is a mere consequence of the complicated nonlinear relationship between the singular vectors and $\mathscr{E}$ induced by the iterations. If the initial values $\mathbf{x}_k^{[0]}$s are independent of $\mathscr{X}$, then running the power iteration (7) once would result in a rate optimal estimate.

Proposition 4.2: Assume that $\mathbf{x}_k^{[0]}$s are independent of $\mathscr{X}$ and satisfy

$$\max_{1\le k\le p} \sin\angle(\mathbf{x}_k^{[0]}, \mathbf{u}_k) \le \eta$$

for some constant $\eta < 1$. Then for any $0 < \delta < 1$,

$$\sin\angle(\mathbf{x}_k^{[1]}, \mathbf{u}_k) \le \min\left\{\frac{C\sqrt{d_k}/\delta^{1/\alpha}}{\lambda(1-\eta^2)^{(p-1)/2}}, 1\right\}$$

for $1 \le k \le p$ with probability at least $1-\delta$.

Proposition 4.2 immediately suggests a simple strategy to estimate $\mathbf{u}_k$s when we observe, in addition to $\mathscr{X}$, another independent copy of it, denoted by $\tilde{\mathscr{X}}$: first apply robust tensor SVD to $\tilde{\mathscr{X}}$, and then update the estimated singular vectors using (7). As a direct consequence of Theorem 4.1 and Proposition 4.2, the resulting estimate $\tilde{\mathbf{u}}_k$s satisfy:

$$\max_{1\le k\le p} \sin\angle(\tilde{\mathbf{u}}_k, \mathbf{u}_k) \lesssim_p \frac{\sqrt{d_k}}{\lambda}. \tag{9}$$

if $\lambda \ge C(\bar{d}_G)^{p/4}\log(\bar{d}_G)$ for a sufficiently large constant $C > 0$.

Of course, we do not have another copy of $\mathscr{X}$. To overcome this obstacle, we randomly partition the tensor into two halves along its k-th mode. Denote the two halves of indices by J_1 and J_2. We use the tensor $\mathscr{X}_1$, with indices $[d_1]\times\ldots[d_{k-1}]\times J_1\times[d_{k+1}]\times[d_p]$ for nontrivial initialization, and $\mathscr{X}_2$ with indices J_1, replaced by J_2 for iteration. It can be derived from the scaled Chernoff bound that

$$\mathbb{P}\left(\sum_{i\in J_1} u_{ki}^2 \ge 0.25\right) \le \exp\left(-1/16\mu_2^2\right).$$

See, e.g., Theorems 1, 2 and the subsequent remarks of [40]. Note that we can write

$$\mathscr{X}_1 = \lambda\left\|\mathbf{u}_{k,J_1}\right\|\mathbf{u}_1\circ\cdots\circ\mathbf{u}_{k-1}\circ\frac{\mathbf{u}_{k,J_1}}{\left\|\mathbf{u}_{k,J_1}\right\|\cdots\circ\mathbf{u}_p} + \mathscr{E}_1.$$

The last two equations imply that $\mathscr{X}_1$ has a signal strength of at least

$$0.5\lambda > C\left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)(\log \bar{d}_G)^{1/4}.$$

Thus we can use Theorem 4.1, assuming all the incoherence conditions are satisfied, to get estimates $\widehat{\mathbf{v}}_k$ such that

$$\max_{1\le k\le p-1} \sin\angle\left(\widehat{\mathbf{v}}_k,\, \mathbf{u}_k\right) \le \eta, \tag{10}$$

for some constant $\eta < 1$. Notice that $\widehat{\mathbf{v}}_k$ are independent of $\mathscr{X}_2$. Following (7), we can use $\mathbf{x}_t^{[0]} = \widehat{\mathbf{v}}_k$ with the second tensor $\mathscr{X}_2$ to yield an improved estimate of $\mathbf{u}_k$, denoted by $\widehat{\mathbf{u}}_k$.

In light of Theorem 4.1 and Proposition 4.2, we get the following theorem.

Theorem 4.3: Assume that the entries of $\mathscr{E}$ are independent and identically distributed with zero mean, unit variance and $\mathbb{E}|E|^\alpha < \infty$ for some $\alpha \ge 2$. There exist constants $C_1, C_2, C_3 > 0$ such that if $\lambda > C_1\left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)(\log \bar{d}_G)^{1/4}$ and $\max_{1\le k\le p}\|\mathbf{u}_k\|_{\ell_\infty} \le C_2(\log \bar{d}_G)^{-1}$, then there is a polynomial time computable estimate $\widehat{\mathbf{u}}_k$ $(k = 1, \ldots, p)$ obeying

$$\mathbb{P}\left\{\sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) \le \frac{C_3\sqrt{d_k}}{\lambda t} \quad \text{for } 1 \le k \le p\right\} \ge 1 - t^\alpha$$

for any $0 < t < 1$. Here $d_{\max}$ and $\bar{d}_G$ are as defined in (3).

Note that the additional requirement of $\max_{1\le k\le p}\|\mathbf{u}_k\|_{\ell_\infty} \le C_2(\log \bar{d}_G)^{-1}$ ensures that the singular vectors are not too concentrated on a few coordinates and therefore allows us to capture the signal even after the sample splitting.

In the event that this is not the case, we provide some heuristic justification about how our task can be effectively reduced to a problem of lower order. To see this, assume, without loss of generality, that $u_{11} = \|\mathbf{u}_1\|_{\ell_\infty} \gtrsim (\log \bar{d}_G)^{-1}$. Denote by $\mathscr{X}_j$ the jth slice of $\mathscr{X}$ along its first mode. It is clear that

$$\mathscr{X}_1 = \tilde{\lambda}\mathbf{u}_2\circ\cdots\circ\mathbf{u}_p + \mathscr{E}_1,$$

where

$$\tilde{\lambda} = \lambda u_{11} \gtrsim \left(d_{\max}^{1/2} + \left(\bar{d}_G\right)^{p/4}\right)\operatorname{polylog}(\bar{d}_G),$$

by assumption. Note that the signal strengths λ and $\tilde{\lambda}$ are of the same order up to the logarithmic factor. However, $\mathscr{X}$ is a pth order tensor and $\mathscr{X}_1$ is of order $(p-1)$.

We can then proceed to decompose the $(p-1)$ order tensor. If the maximum entry $\|\mathbf{u}_k\|_{\ell_\infty}$ for $k \ge 2$, is smaller than $(\log \bar{d}_G)$, we can apply Theorem 4.1 to obtain estimates. If not, we repeat the procedure described above, now with $\mathbf{u}_2$, to get a $(p-2)$ order tensor. Since the tensor order decreases at each step, estimating the singular vectors could become successively easier because of the relative higher signal-to-noise ratio.

Finally, notice that the robust estimation method of the present section does not depend on α, provided $\alpha \ge 2$. This allows the user to apply this method without any prior knowledge about the error distribution. The numerical experiments of Section 5 also support this claim. When the signal strength condition is satisfied, the performance of the robust estimators does not depend on the number of moments of the errors.

V. Numerical Experiments

To complement the theoretical developments, we also conducted several sets of numerical experiments. In all the simulations, we set $d_1 = d_2 = d_3 := d$. In the first set, we take $d = 400$, $\mathscr{T} = \lambda\mathbf{u}_1\circ\mathbf{u}_2\circ\mathbf{u}_3 + \mathscr{E}$, where $\lambda = 3d^{3/4}$ and

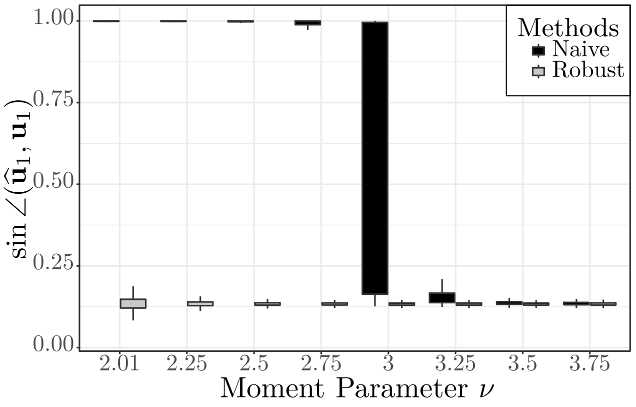

Fig. 4. Comparison of the methods for Pareto(ν) errors across different values of ν. For each ν, the black bar on the left corresponds to the näive estimate based on HOSVD, the gray bar on the right refers to the robust tensor SVD.

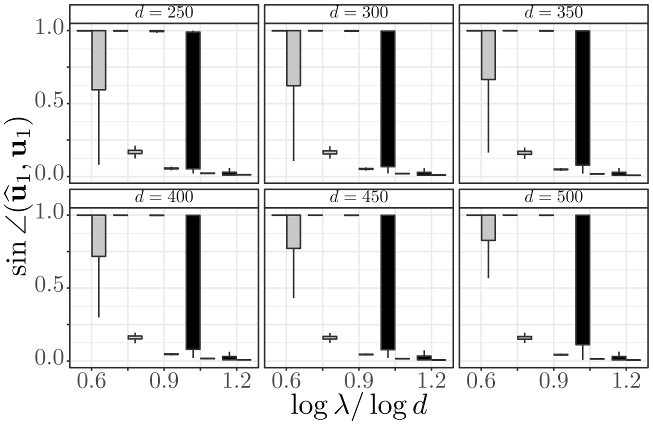

Fig. 5. Comparison of the methods for Pareto(2.1) errors across dimension. In each panel, for each value of $\log \lambda / \log d$, the black bar on the left corresponds to the naïve estimate based on HOSVD, the gray bar on the right corresponds to the robust tensor SVD.

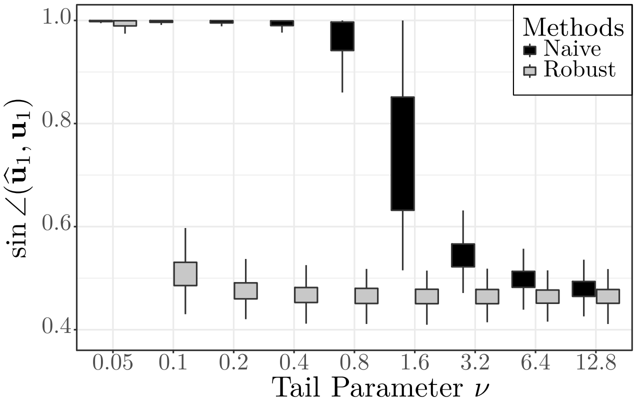

Fig. 6. Comparison of methods for mixture distributed errors across a tail parameter ν.

$\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ were sampled uniformly from the unit sphere. The elements of $\mathscr{E}$ are independently simulated from symmetrized and appropriately scaled Pareto distributions. More specifically, we generated $E_{ijk} = P_{ijk}R_{ijk}/\sqrt{\nu/(\nu-2)}$, where $P_{ijk} \sim \text{Pareto}(\nu)$ and R_{ijk}s are i.i.d. Rademacher random variables. The rescaling was done to ensure the errors have unit variance. Note that E_{ijk} has finite αth moment if and only if $\nu > \alpha$. We therefore varied ν to simulate noises satisfying different moment conditions. We ran the algorithm in Section IV with an initial guess of 3000 for λ. Even though this is a huge overestimate, it does not affect the final results. For comparison, we also computed the naïve estimate based on HOSVD. The results from 1000 simulation runs for each value of ν are summarized in Figure 4. It can be observed that the robust method provides an estimate that is strongly correlated with the true vector $\mathbf{u}_1$, irrespective of ν. On the other hand, the naïve estimate is almost orthogonal to the signal direction for smaller values of ν, but its performance improves as ν approaches 4, as predicted by Proposition 3.1.

We next provide a numerical experiment to corroborate the signal strength requirements for consistent estimation. The setup is similar to before and we fixed $\nu = 2.1$ and varied d from 250 to 500. We took $\lambda = 3d^{\xi}$ for $\xi = 0.6, 0.75, \ldots, 1.2$ to correspond to different signal strength. The result, again summarized from 1000 simulation runs, is presented in Figure 5. It indicates that $\xi = 3/4$ is indeed the correct computational threshold. When $\xi < 0.75$, neither of the methods is successful. However, as soon as ξ reaches 0.75, the robust SVD method from Section IV is able to provide nontrivial estimates. The accuracy improves as ξ increases further. On the other hand, the naïve estimator performs poorly for ξ as large as 1.05, where it has a very large variance, before transitioning to a better estimate at $\xi = 1.2$.

To investigate the possible effect of different error distributions or lack thereof, we also considered a simulation setting similar to the one used by [28]. We fixed $d = 400$ and set $\mathscr{T} = \lambda \mathbf{u}_1 \circ \mathbf{u}_2 \circ \mathbf{u}_3 + \mathscr{E}$, where $\lambda = 1.5d^{3/4}$ and $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are sampled uniformly from the unit sphere. The errors are independently distributed as $R_{ijk}X_{ijk}$ where R_{ijk}s are Rademacher random variables while $X_{ijk} = -\sqrt{\dfrac{d-\nu}{\nu}}$ with probability $\dfrac{\nu}{d}$ and $X_{ijk} = \sqrt{\dfrac{\nu}{d-\nu}}$ with probability $1 - \dfrac{\nu}{d}$. The distribution becomes lighter tailed as ν increases. The robust method still has better performance than the naïve one, even for much lighter tailed errors. We arbitrarily fixed the truncation parameter $\theta = 0.2$ and used a single robust iteration with no sample splitting. As shown by [28], this error distribution can worsen the performance of elementwise truncation, however our experiment results, summarized from 1000 simulations in Figure 6, confirms that this has no effect on the spectrum truncated estimators that we proposed.

We also examined the effect of signal strength for this noise distribution. We fixed the mixture parameter $\nu = 0.1$ and vary the dimension d from 200 to 450, while setting $\lambda = 1.5\ d^{\xi}$. The results summarized from 1000 simulations is given in Figure 7. The observation is similar to before: the robust SVD method is successful whenever $\xi \geq 0.75$. The naïve estimator is almost orthogonal to the signal till $\xi = 0.8$, then goes through a high variance phase at $\xi = 0.85$, finally providing a nontrivial estimate when $\xi = 0.9$.

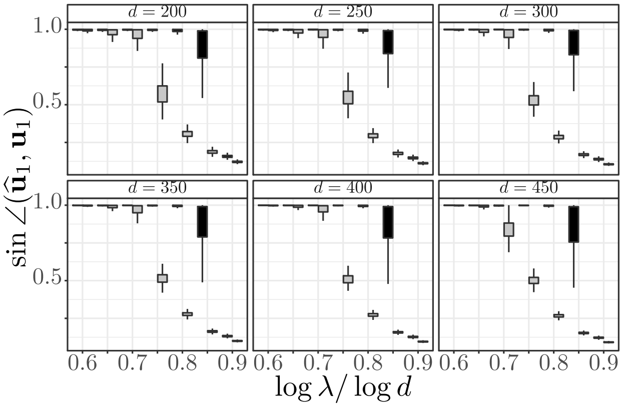

Fig. 7. Comparison of the methods for mixture distributed errors across dimension. In each panel, for each value of $\log\lambda/\log d$, the black bar on the left corresponds to the naïve estimate based on HOSVD, the gray bar on the right corresponds to the robust tensor SVD.

VI. Concluding Remarks

In this paper, we study the problem of estimating the rank-one spikes in the presence of heavy tailed noises. Our contributions are three-fold. First, we investigate the performance of estimates from tensor SVD, perhaps the most natural approach especially if we neglect the computational cost. Our results identify the signal strength requirement for the tensor SVD to yield rate-optimal estimates. (Nearly) matching lower bounds are also given to show that these requirements are optimal in the sense that the tensor SVD is necessarily inconsistent if the signal strength is below them.

Our analysis of the tensor SVD exploits its close connection with the spectral norm of random tensors, and our second contribution is to establish upper bounds and (nearly) matching lower bounds for a tensor consisting of independent mean zero random variables. Our bounds pinpoint the connection between spectral norm of a random tensor and the moment condition for its entries.

Finally, we develop procedures for estimating the singular vectors under heavy tailed noises that are tractable in that they are polynomial time computable, practical in that they are easy to implement, and yields estimates that converge to the true parameter at the optimal rate. In particular, we show that similar to the case with Gaussian noise, a single power iteration with spectral initialization suffices if the entries of the noise have finite $4(p-1)$th moment. If the entries have finite fourth moment but infinite $4(p-1)$th moment, then we need to do $O(\log d_{\max})$ number of power iterations. If the entries do not have finite fourth moment, we need a different strategy. This new procedure combines robust matrix estimation and sample splitting, and can be shown as both tractable and rate optimal.

Our work also points to a number of interesting directions that warrant further investigation. For example, one may consider the effect of heavy tail noise beyond the rank one case. The rank-one spike model we considered here can be generalized along two directions: the Tucker decomposition framework (see e.g., [17]), or through orthogonally decomposable tensors (see e.g., [18]). It is immediate that our results continue to hold in both these frameworks when the rank is at most a constant, i.e., does not grow with $d_{\min}$. A more careful inspection shows that the matricization/power iteration estimator considered in Section III, will still succeed for Tucker decomposition for any sequence of multiranks if the errors have finite fourth moments. The case where $2 < \alpha < 4$ is more nuanced. When all the spikes are of different magnitude, we can adapt a successive rank one method to compute the leading singular vector at each level. However, more interestingly, such low rank tensors have a unique decomposition even when there are $r > 1$ orthogonal spikes of the same magnitude. See, e.g., [18] for further discussion. It is plausible that a more sophisticated truncation strategy would be required to recover the unique decomposition in that case, especially when $r/d \to c > 0$. We leave this intriguing question for future study.

Finally, our numerical experiments suggest that it is possible that there are sharper bounds on the critical value of λ, depending on specific error distributions. For example, in Figure 4, while the moment dependent error bounds are sharp in the end cases where the moment parameters are $\nu = 2$ and $\nu = 4$, the naive matricization estimator seems to pass through a high variance phase at $\nu = 3$. A Monte Carlo calculation shows that in that case, the error is smaller than the specified signal value (and hence the naive estimator is successful) with a nontrivial probability. Deriving more precise bounds in this subcritical regime is of clear interest.

VII. Proofs

Throughout this section we will write C to mean a constant that may differ from line to line. Similarly we write C_p to refer to a constant that depends on the tensor order p.

A. Moment Bounds for Random Tensors

The proof of Theorems 2.1 and 2.2 uses Talagrand's concentration inequality for convex Lipschitz functions combined with estimates of higher order moments via Khintchine and Rosenthal inequalities. In particular, it relies on the following moment bound for random tensors which may be of independent interest.

Theorem 7.1: Let $\mathscr{E} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ be a pth order random tensor whose entries are independent such that $\mathbb{E}E_{i_1\ldots i_p} = 0$ and $\mathbb{E}E^2_{i_1\ldots i_p} = \sigma^2_{i_1\ldots i_p}$. Then for any $r \geq 1$, there is a constant C_p depending only on p such that

$$\begin{aligned}
&(\mathbb{E}\,\|\mathscr{E}\|^r)^{\frac{1}{r}} \\
&\leq C_p\sigma\sqrt{d_{\max}} \\
&\quad + C_p\sigma^{\frac{1}{2}}(\log d_{\max})^{\frac{3}{4}}\times \\
&\quad\times\left(\sum_{k=1}^{p} d_k\mathbb{E}\left(\max_{i_l\in[d_l],l\neq k}\left|\sum_{i_k=1}^{d_k}(E^2_{i_1\ldots i_p}-\sigma^2_{i_1\ldots i_p})\right|\right)^{\frac{r}{4}}\right)^{\frac{1}{r}} \\
&\quad + C_p(\log d_{\max})\times \\
&\quad\times\left(\sum_{i_k=1}^{p}\mathbb{E}\max_{i_l\in[d_l],l\neq k}\left|\sum_{i_k=1}^{d_k}(E^2_{i_1\ldots i_p}-\sigma^2_{i_1\ldots i_p})\right|^{r/2}\right)^{\frac{1}{r}},
\end{aligned}$$

where $\sigma^2 = \max_{i_1,\ldots,i_p}\sigma^2_{i_1\ldots i_p}$.

We want to point out that this theorem is a stronger version of the results derived by [38]. In particular, we reduce the $\log d_{\max}$ factor in all cases, to the point that it does not depend on p, the order of the tensor. More importantly, the leading term in our error bound is free of the $\log d_{\max}$ factor. This allows us to derive bounds differing only by a constant, whenever $\alpha > 4(p-1)$, a conclusion which is in tune with the corresponding results for matrices. The same conclusion cannot be drawn from Theorem 2 and Corollary 2 of [38].

We want to mention here that for $p = 2$, i.e., matrices, these bounds can be compared to the matrix norm bounds from [25]. Our bounds are worse only by the $(\log d_{\max})$ factors.

Note that we do not assume that the entries of $\mathscr{E}$ are identically distributed in Theorem 7.1. In fact, it follows directly that the upper bounds in Theorems 2.1 and 2.2 continue to hold if we have independent, but not necessarily identically distributed errors, as long as the moment conditions are satisfied. We opt for the current version of Theorems 2.1 and 2.2 for ease of exposition.

It is not hard to see that

$$\|\mathscr{E}\| \geq \max_{k\in[p]} \max_{i_l \in [d_l], l\neq k} \left(\sum_{i_k=1}^{d_k} E_{i_1\ldots i_p}^2 \right)^{1/2}.$$

This immediately suggests that

$$\left(\mathbb{E}\,\|\mathscr{E}\|^r\right)^{\frac{1}{r}} \gtrsim \left(\sum_{k=1}^{p} \mathbb{E} \left(\max_{i_l\in[d], l\neq k} \sum_{i_k=1}^{d_k} E_{i_1\ldots i_p}^2 \right)^{\frac{r}{2}} \right)^{\frac{1}{r}}.$$

The lower bound above matches the upper bound in Theorem 7.1 up to the $\log d$ terms for any fixed p. Indeed, a close inspection of the proof of Theorem 7.1 indicates that the $\log d$ terms in the upper bound may be removed altogether with some stronger moment assumptions. The proof of Theorem 7.1 relies on a scheme developed earlier by [41] and is similar in spirit to that from [38].

Proof of Theorem 7.1: By the standard symmetrization argument and conditioning (see, e.g., Lemma 5 of [38]),

$$\left(\mathbb{E}(\|\mathscr{E}\|)^r\right)^{1/r} \leq \sqrt{2\pi}\,\mathbb{E}_{\mathscr{E}} \left(\mathbb{E} \left(\|\mathscr{H}\|^r \,\middle|\, \mathscr{E} \right) \right)^{1/r}, \tag{11}$$

where $\mathscr{H}$ is a $d_1 \times \cdots \times d_p$ tensor with entries $H_{i_1\ldots i_p} = E_{i_1\ldots i_p} Z_{i_1\ldots i_p}$, $Z_{i_1\ldots i_p} \overset{iid}{\sim} N(0,1)$. Let $\sigma_{i_1\ldots i_p}^2 = \mathbb{E}E_{i_1\ldots i_p}^2$ and

$$\sigma^2 := \max_{i_1\ldots i_p} \sigma_{i_1\ldots i_p}^2.$$

We will first show that for any fixed tensor $\mathscr{E}$, $\mathscr{H}$ defined above satisfies the following inequality.

$$\begin{aligned}
&\left(\mathbb{E}\,\|\mathscr{H}\|^r\right)^{1/r} \\
\leq &C_p \sigma \sqrt{d_{\max}} \\
&+ C_p \sigma^{\frac{1}{2}} (\log d_{\max})^{\frac{3}{4}} \times \\
&\times \left(\sum_{k=1}^{p} d_k \left(\max_{i_l\in[d_l], l\neq k} \left| \sum_{i_k=1}^{d_k} (E_{i_1\ldots i_p}^2 - \sigma_{i_1\ldots i_p}^2) \right| \right)^{\frac{r}{4}} \right)^{1/r} \\
&+ C_p (\log d_{\max}) \times \\
&\times \left(\sum_{i_k=1}^{p} \max_{i_l\in[d_l], l\neq k} \left| \sum_{i_k=1}^{d_k} (E_{i_1\ldots i_p}^2 - \sigma_{i_1\ldots i_p}^2) \right|^{r/2} \right)^{1/r}.
\end{aligned} \tag{12}$$

To this end, we shall use an ϵ-net argument.

For any integer L, write $S_L = \{0, 1, \ldots, 2^{-L}\}$. It follows from Lemma 3 that the set $N_{L_q}^{(q)} = \{\mathbf{x} \in \mathbb{R}^{d_q} : \|\mathbf{x}\| \leq 1, x_i^2 \in S_L\}$ forms a $(1/2)$-net for $\mathbb{S}^{d_q-1}$, $1 \leq q \leq p$ by taking $L_q = \log d_q + c_0$ for some constant c_0. Now, for $\mathbf{x} \in \mathbb{R}^{d_q}$, define the projections $\Pi_l(\mathbf{x})$, $\Pi_{<l}(\mathbf{x}) \in \mathbb{R}^{d_q}$ s.t.

$$(\Pi_l(\mathbf{x}))_i = x_i \mathbb{1}(x_i^2 = 2^{-l})$$
$$(\Pi_{<l}(\mathbf{x}))_i = x_i \mathbb{1}(x_i^2 \geq 2^{-l})$$

for $1 \leq i \leq d_q$. Let $L := \max_{1\leq q\leq p} L_q$. Let us also define $N_l^{(q)} = \Pi_l(N_{L_q}^{(q)})$ and $N_{<l}^{(q)} = \Pi_{<l}(N_{L_q}^{(q)})$ for $1 \leq l \leq L$. We have, for any $\mathbf{x} \in N_{L_q}^{(q)}$,

$$\mathbf{x} = \sum_{l=1}^{L} \Pi_l(\mathbf{x}) \qquad \text{and} \qquad \sum_{m<l} \Pi_m(\mathbf{x}) = \Pi_{<l}(\mathbf{x}).$$

Note that if $L_q < L$ for some $1 \leq q \leq p$, we have $\Pi_l(\mathbf{x}) = 0$ for all $L_q < l \leq L$ and all $\mathbf{x} \in N_{L_q}^{(q)}$. Expanding the sum for each vector $\mathbf{x}_q \in \mathbb{R}^{d_q}$, we get

$$\begin{aligned}
&\mathscr{H} \times_2 \mathbf{x}_2 \cdots \times_p \mathbf{x}_p \\
&= \sum_{l_1=1}^{L} \cdots \sum_{l_p=1}^{L} \mathscr{H} \times_2 \Pi_{l_2}(\mathbf{x}_2) \cdots \times_p \Pi_{l_p}(\mathbf{x}_p) \\
{}^{*}&= \sum_{k=2}^{p} \sum_{l_k=1}^{L} \sum_{\substack{l_i \leq l_k \\ i\neq k}} \mathscr{H} \times_2 \Pi_{l_2}(\mathbf{x}_2) \cdots \times_p \Pi_{l_p}(\mathbf{x}_p) \\
&= \sum_{k=2}^{p} \sum_{l_k=1}^{L} \mathscr{H} \times_2 \left(\sum_{l_2\leq l_k} \Pi_{l_2}(\mathbf{x}_2) \right) \cdots \times_k (\Pi_{l_k}(\mathbf{x}_k)) \\
&\qquad \times_{k+1} \cdots \times_p \left(\sum_{l_p \leq l_k} \Pi_{l_p}(\mathbf{x}_p) \right) \\
&= \sum_{k=2}^{p} \sum_{l=1}^{L} \mathscr{H} \times_2 \Pi_{<l}(\mathbf{x}_2) \cdots \times_{k-1} \Pi_{<l}(\mathbf{x}_{k-1}) \times_k \Pi_l(\mathbf{x}_k) \\
&\qquad \times_{k+1} \cdots \times_p \Pi_{<l}(\mathbf{x}_p).
\end{aligned}$$

By triangle inequality,

$$\begin{aligned}
&\|\mathscr{H}\|^2 = \sup_{\mathbf{x}_q\in\mathbb{S}^{d_q-1},\, 2\leq q\leq p} \|\mathscr{H} \times_2 \mathbf{x}_2 \cdots \times_p \mathbf{x}_p\|^2 \\
&\leq 2^{2p-2} \max_{\mathbf{x}_q \in N_{L_q}^{(q)}, 2\leq q\leq p} \|\mathscr{H} \times_2 \mathbf{x}_2 \cdots \times_p \mathbf{x}_p\|^2 \\
&\leq 2^{2p-2} \times \\
&\times \max_{\substack{\mathbf{x}_q \in N_{L_q}^{(q)} \\ 2\leq q\leq p}} \left[\sum_{k=2}^{p} \left\| \sum_{l=1}^{L} \mathscr{H} \times_2 \Pi_{<l}(\mathbf{x}_2) \cdots \times_{k-1} \Pi_{<l}(\mathbf{x}_{k-1}) \right.\right. \\
&\qquad \left.\left. \times_k \Pi_l(\mathbf{x}_k) \times_{k+1} \cdots \times_p \Pi_{<l}(\mathbf{x}_p) \right\| \right]^2
\end{aligned}$$

[*] Let $l_* = \max l_q$, then we take $k = \min\{1 \leq q \leq p : l_q = l_*\}$.

$$\begin{aligned}&\leq 2^{2p-2}p\times\\ &\times\sum_{k=2}^{p}\max_{\substack{\mathbf{x}_q\in N^{(q)}_{L_q}\\ 2\leq q\leq p}}\left\|\sum_{l=1}^{L}\mathscr{H}\times_2\Pi_{<l}(\mathbf{x}_2)\cdots\times_{k-1}\Pi_{<l}(\mathbf{x}_{k-1})\right.\\ &\qquad\qquad\left.\times_k\Pi_l(\mathbf{x}_k)\times_{k+1}\cdots\times_p\Pi_{<l}(\mathbf{x}_p)\right\|^2. \qquad (13)\end{aligned}$$

Without loss of generality, we assume $d_2 \geq d_3 \geq \cdots \geq d_p$. Because of symmetry, we shall focus on $k = 2$ without loss of generality. To simplify notation, let us denote

$$\mathbf{T}_l(\mathbf{x}_1,\ldots,\mathbf{x}_p) = \mathscr{H}\times_2\Pi_l(\mathbf{x}_2)\times_3\Pi_{<l}(\mathbf{x}_3)\cdots\times_p\Pi_{<l}(\mathbf{x}_p).$$

For any fixed $\mathbf{x}_2,\ldots,\mathbf{x}_p$, we have

$$\begin{aligned}&\left\|\sum_{l=1}^{L}\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p)\right\|^2\\ &=\sum_{l=1}^{L}\|\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p)\|^2\\ &\quad+2\sum_{l<l'}(\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p))^\top\mathbf{T}_{l'}(\mathbf{x}_2,\ldots,\mathbf{x}_p)\\ &=\sum_{l=1}^{L}\sum_{i_1=1}^{d_1}(\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p))_{i_1}^2+\\ &\quad+\sum_{l=1}^{L}\sum_{i_1=1}^{d_1}(\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p))_{i_1}(\mathbf{U}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p))_{i_1} \qquad (14)\end{aligned}$$

where we define $\mathbf{U}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p) = \sum_{l'>l}\mathbf{T}_{l'}(\mathbf{x}_2,\ldots,\mathbf{x}_p)$. We bound the sum of squares term first. To reduce notation, here onward we will refer to $\mathbf{T}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p)$, $\mathbf{U}_l(\mathbf{x}_2,\ldots,\mathbf{x}_p)$ as $\mathbf{T}_l(\mathbf{x})$, $\mathbf{U}_l(\mathbf{x})$ or simply $\mathbf{T}_l$, $\mathbf{U}_l$ respectively when there is no scope of confusion.

Note that conditional on $\mathscr{E}$, one can show that $\mathbf{T}_l \sim N\left(\mathbf{0}, \operatorname{diag}(\tau_{1l}^2,\ldots,\tau_{d_1l}^2)\right)$, where

$$\tau_{i_1l}^2=\sum_{i_2=1}^{d_2}\cdots\sum_{i_p=1}^{d_p}E_{i_1\ldots i_p}^2\Pi_l(\mathbf{x}_2)_{i_2}^2\ldots\Pi_{<l}(\mathbf{x}_p)_{i_p}^2. \qquad (15)$$

In particular, the co-ordinates of $\mathbf{T}_l$, given $\mathscr{E}$, are independent, $(\mathbf{T}_l)_{i_1} \sim N(0,\tau_{i_1l}^2)$ for $1 \leq i_1 \leq d_1$.

Conditional on $\mathscr{E}$, we use the Bernstein inequality for sub-exponential random variables (see, e.g., Theorem 2.8.1 of [42]), to obtain

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i_1=1}^{d_1}\left((\mathbf{T}_l)_{i_1}^2-\tau_{i_1l}^2\right)\geq C\sqrt{t\sum_{i_1}\tau_{i_1l}^4}+Ct\max_{i_1}\tau_{i_1l}^2|\mathscr{E}\right\}\\ &\leq\exp(-t).\end{aligned}$$

Here onward, all probability statements are conditional on $\mathscr{E}$, unless otherwise mentioned. For notational convenience, we avoid repeating this statement. By Lemma 4, for $1 \leq q \leq p$,

$$|N_l^{(q)}| < |N_{<l}^{(q)}| < \exp(C2^l(1+L-l)).$$

An application of the union bound yields

$$\begin{aligned}&\bigcup_{\mathbf{x}_q\in N_l^{(q)},2\leq q\leq p}\|\mathbf{T}_l\|^2\\ &\geq\sum_{i_1=1}^{d_1}\tau_{i_1l}^2+C\sqrt{((p-1)2^l(1+L-l)+t)\sum_{i_1}\tau_{i_1}^4}\\ &\quad+C((p-1)2^l(1+L-l)+t)\max_{i_1}\tau_{i_1l}^2\end{aligned}$$

with probability at most $\exp(-t)$. Summing over l for each fixed $\mathbf{x}_j$, and by union bound over l,

$$\begin{aligned}&\bigcup_{\mathbf{x}_q\in N_{L_q}^{(q)},2\leq q\leq p}\sum_{l=1}^{L}\|\mathbf{T}_l\|^2\\ &\geq\sum_{l=1}^{L}\sum_{i_1=1}^{d_1}\tau_{i_1l}^2+C\sum_{l=1}^{L}\sqrt{((p-1)2^l(1+L-l)+t)\sum_{i_1}\tau_{i_1l}^4}\\ &\quad+C\sum_{l=1}^{L}((p-1)2^l(1+L-l)+t)\max_{i_1}\tau_{i_1l}^2 \qquad (16)\end{aligned}$$

with probability at most $L\exp(-t)$.

The cross-product term can be bounded similarly. Note that $\mathbf{U}_l \sim N(\mathbf{0}, \operatorname{diag}(\gamma_{1l}^2,\ldots,\gamma_{d_1}^2))$ where $\gamma_{i_1l}^2 = \sum_{l'>l}\tau_{i_1l'}^2$ for $1 \leq i_1 \leq d_1$. Once again, $(\mathbf{T}_l)_{i_1}(\mathbf{U}_l)_{i_1}$ are independent and satisfy

$$\mathbb{E}(\mathbf{T}_l)_{i_1}(\mathbf{U}_l)_{i_1}=0,\quad\|(\mathbf{T}_l)_{i_1}(\mathbf{U}_l)_{i_1}\|_{\psi_1}\leq\tau_{i_1l}\gamma_{i_1l},$$

where $\|\cdot\|_{\psi_1}$ is the subexponential norm. Then similarly using Bernstein inequality followed by a union bound over the special ϵ-net, we have, just as in (16), that

$$\begin{aligned}&\bigcup_{\mathbf{x}_q\in N_{L_q}^{(q)},2\leq q\leq p}\sum_{l=1}^{L}\mathbf{T}_l^\top\mathbf{U}_l\\ &\geq C\sum_{l=1}^{L}\sqrt{((p-1)2^l(1+L-l)+t)\sum_{i_1}\tau_{i_1l}^2\gamma_{i_1l}^2}\\ &\quad+C\sum_{l=1}^{L}((p-1)2^l(1+L-l)+t)\max_{i_1}\tau_{i_1l}\gamma_{i_1l} \qquad (17)\end{aligned}$$

with probability at most $L\exp(-t)$.

We bound the "sum of expectations" term in (16) as

$$\begin{aligned}&\sum_{l=1}^{L}\sum_{i_1=1}^{d_1}\tau_{i_1l}^2\\ &\leq\sum_{l=1}^{L}\sum_{i_2=1}^{d_2}\Pi_l(\mathbf{x}_2)_{i_2}^2\sum_{i_3\ldots i_p}\prod_{j=3}^{p}\Pi_{<l}(\mathbf{x}_j)_{i_j}^2\left(\sum_{i_1=1}^{d_1}E_{i_1\ldots i_p}^2\right)\\ &\leq\sum_{l=1}^{L}\sum_{i_2=1}^{d_2}\Pi_l(\mathbf{x}_2)_{i_2}^2\max_{i_3,\ldots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\ldots i_p}^2\\ &\leq\max_{i_2,\ldots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\ldots i_p}^2\sum_{i_2=1}^{d_2}\sum_{l=1}^{L}\Pi_l(\mathbf{x}_2)_{i_2}^2\\ &\leq\max_{i_2,\ldots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\ldots i_p}^2\sum_{i_2=1}^{d_2}\mathbf{x}_{2i_2}^2\\ &\leq\max_{i_2,\ldots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\ldots i_p}^2. \qquad (18)\end{aligned}$$

It remains to bound the deviation terms in both (16) and (17). To that end, notice that, for any fixed $i_1 \in [d_1]$, by (15), since $\|\Pi_{<l}(\mathbf{x}_q)\| \leq 1$ we have

$$\begin{aligned}
&\gamma_{i_1 l}^2 \\
&= \sum_{l'>l} \tau_{i_1 l'}^2 \\
&\leq \sum_{i_3,\dots,i_p} \prod_{j=3}^{p} \Pi_{<l'}(\mathbf{x}_j)_{i_j}^2 \sum_{l'>l} \sum_{i_2=1}^{d_2} \Pi_{l'}(\mathbf{x}_2)_{i_2}^2 E_{i_1\dots i_p}^2 \\
&\leq \max_{i_3,\dots,i_p} \sum_{l'>l} \sum_{i_2=1}^{d_2} \Pi_{l'}(\mathbf{x}_2)_{i_2}^2 E_{i_1\dots i_p}^2 \\
&\leq 2^{-l} \max_{i_3,\dots,i_p} \sum_{i_2=1}^{d_2} E_{i_1\dots i_p}^2 \mathbb{1}(\Pi_{l'}(\mathbf{x}_2)_{i_2} \neq 0 \text{ for some } l' > l) \\
&\leq \sigma^2 + 2^{-l} \max_{i_3,\dots,i_p} \sum_{i_2=1}^{d_2} \Big[(E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \\
&\qquad \mathbb{1}(\Pi_{l'}(\mathbf{x}_2)_{i_2} \neq 0 \text{ for some } l' > l) \Big]
\end{aligned} \tag{19}$$

where we use the definition $\sigma^2 := \max \sigma_{i_1\dots i_p}^2$ and

$$\sum_{i_2} \mathbb{1}(\Pi_{l'}(\mathbf{x}_2)_{i_2} \neq 0 \text{ for some } l' > l) \leq \sum_{l'>l} 2^{-l'} = 2^{-l}.$$

We similarly have

$$\begin{aligned}
&\tau_{i_1 l}^2 \\
&\leq 2^{-l} \max_{i_3\dots i_p} \sum_{i_2=1}^{d_2} E_{i_1\dots i_p}^2 \mathbb{1}(\Pi_l(\mathbf{x}_2)_{i_2} \neq 0) \\
&\leq \sigma^2 + 2^{-l} \max_{i_3\dots i_p} \sum_{i_2=1}^{d_2} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \mathbb{1}(\Pi_l(\mathbf{x}_2)_{i_2} \neq 0).
\end{aligned} \tag{20}$$

and the following inequality.

$$\begin{aligned}
&\sum_{i_1} \tau_{i_1 l}^2 \\
&\leq \sum_{i_3\dots i_p} \prod_{j=3}^{p} \Pi_{<l'}(\mathbf{x}_j)_{i_j}^2 \sum_{l'>l} \sum_{i_2=1}^{d_2} \Pi_{l'}(\mathbf{x}_2)_{i_2}^2 \sum_{i_1} E_{i_1\dots i_p}^2 \\
&\leq 2^{-l} \max_{i_3,\dots,i_p} \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} E_{i_1\dots i_p}^2 \mathbb{1}(\Pi_l(\mathbf{x}_2)_{i_2} \neq 0) \\
&\leq d_1 \sigma^2 \\
&\quad + \max_{i_2 i_3\dots i_p} \left| \sum_{i_1=1}^{d_1} E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2 \right| 2^{-l} \sum_{i_2=1}^{d_2} \mathbb{1}(\Pi_l(\mathbf{x}_2)_{i_2} \neq 0).
\end{aligned} \tag{21}$$

We define $t_l = 2^{-l} t/(p-1)$. The l^{th} terms in the "sub-Gaussian" deviation term in (16) and (17) can now be bounded as

$$\begin{aligned}
&2^l (1 + L - l + t_l) \sum_{i_1} \tau_{i_1 l}^2 \max\{\gamma_{i_1 l}^2, \tau_{i_1 l}^2\} \\
&\leq 2^l (1 + L - l + t_l) \max_{i_1} \max\{\gamma_{i_1 l}^2, \tau_{i_1 l}^2\} \sum_{i_1} \tau_{i_1 l}^2 \\
&\leq 2^l (1 + L - l + t_l) d_1 \sigma^4 \\
&\quad + (1 + L - l) d_1 \sigma^2 \cdot \max_{i_1, i_3, \dots, i_p} \left| \sum_{i_2=1}^{d_2} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right| \\
&\quad + (1 + L - l + t_l) \sigma^2 \max_{i_2, i_3, \dots, i_p} \left| \sum_{i_1=1}^{d_1} E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2 \right| \\
&\quad + (1 + L - l + t_l) \max_{i_1, i_3 \dots, i_p} \left| \sum_{i_2=1}^{d_2} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right|^2 \\
&\quad + (1 + L - l + t_l) \max_{i_2, i_3 \dots, i_p} \left| \sum_{i_1=1}^{d_1} E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2 \right|^2
\end{aligned}$$

where we use the AM-GM inequality $ab \leq (a^2 + b^2)$ in the last step. In the above we have used the facts that $\|*\| \, \Pi_{<l}(\mathbf{x}_j) \leq 1$ and $\mathbf{x}_2 = \sum \Pi_l(\mathbf{x}_2)$.

For the other terms we will use the facts that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, and moreover

$$\begin{aligned}
\sum_l 2^{l/2} \sqrt{1 + L - l + t_l} &\leq \sum_l 2^{l/2} (1 + L - l) + \frac{\sqrt{t}}{p-1} \\
&\leq 2^{L/2} + \sqrt{t}/(p-1).
\end{aligned}$$

The first part here follows from Lemma 2 with $a = 1/2$. We can also bound

$$\begin{aligned}
\sum_l \sqrt{1 + L - l + t_l} &\leq \sum_l \sqrt{1 + L - l} \\
&\quad + \sum_l 2^{-l/2} \sqrt{t} / \sqrt{p-1} \\
&\leq L^{3/2} + C\sqrt{t}.
\end{aligned}$$

With the above inequalities, one obtains

$$\begin{aligned}
&\sum_{l=1}^{L} \sqrt{2^l (1 + L - l + t_l) \sum_{i_1} \tau_{i_1 l}^2 \max\{\gamma_{i_1 l}^2, \tau_{i_1 l}^2\}} \\
&\leq \sigma^2 \sqrt{d_1} \sum_{l=1}^{L} 2^{l/2} \sqrt{(1 + L - l + t_l)} \\
&\quad + \sigma \sqrt{d_1} (L^{3/2} + C\sqrt{t}) \max_{i_1, i_3, \dots, i_p} \left| \sum_{i_2=1}^{d_2} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right|^{\frac{1}{2}} \\
&\quad + \sigma (L^{3/2} + C\sqrt{t}) \max_{i_2, i_3, \dots, i_p} \left| \sum_{i_1=1}^{d_1} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right|^{1/2} \\
&\quad + (L^{3/2} + C\sqrt{t}) \max_{i_1, i_3, \dots, i_p} \left| \sum_{i_2=1}^{d_2} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right| \\
&\quad + (L^{3/2} + C\sqrt{t}) \max_{i_2, i_3, \dots, i_p} \left| \sum_{i_1=1}^{d_1} (E_{i_1\dots i_p}^2 - \sigma_{i_1\dots i_p}^2) \right|
\end{aligned}$$

$$\begin{aligned}
&\leq C\sigma^2\sqrt{d_1}(2^{L/2}+C\sqrt{t})\\
&\quad+\sigma\sqrt{d_1}(L^{3/2}+C\sqrt{t})\max_{i_1 i_3\dots i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|^{1/2}\\
&\quad+(L^{3/2}+C\sqrt{t})\max_{i_1,i_3,\dots,i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|\\
&\quad+(L^{3/2}+C\sqrt{t})\max_{i_2,i_3,\dots,i_p}\left|\sum_{i_1=1}^{d_1}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|. \quad (22)
\end{aligned}$$

Finally we bound the "sub-exponential" deviation term in (16) and (17) as follows. Once again using equations (19) and (20), we have

$$\begin{aligned}
&\sum_{l=1}^{L}2^l(1+L-l+t_l)\max_{i_l}\max\{\tau_{i_1 l}\gamma_{i_1 l},\tau^2_{i_1 l}\}\\
\leq&\sum_{l=1}^{L}(2^l(1+L-l)+t_l)\max_{i_l}\max\{(\tau^2_{i_1 l}+\gamma_{i_1 l})/2,\tau^2_{i_1 l}\}\\
\leq&\sum_{l=1}^{L}2^l(1+L-l+t_l)\sigma^2\\
&+\sum_{l=1}^{L}(1+L-l+t_l)\left\{\max_{i_1 i_3\dots i_p}\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right\}\\
\leq&\sigma^2\cdot C(2^L+t/(p-1))\\
&+(L^2+t/(p-1))\left\{\max_{i_1 i_3\dots i_p}\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right\},
\end{aligned} \quad (23)$$

once again using $\sum_l 2^l(1+L-l)\leq C\cdot 2^L$ (which follows from Lemma 2 with $a=1$) and $\sum_l t_l=\sum_l 2^{-l}t/(p-1)\leq t/(p-1)$. Plugging in the bounds (18), (22) and (23) into (16) and (17), we have

$$\begin{aligned}
&\bigcup_{\mathbf{x}_q\in N^{(q)}_{L_q},2\leq q\leq p}\sum_{l=1}^{L}\|\mathbf{T}_l\|^2\\
\leq&\max_{i_2,\dots,i_p}\sum_{i_1=1}^{d_1}E^2_{i_1\dots i_p}+\sigma^2C(d_{\max}+\sqrt{d_1d_{\max}}+t+\sqrt{d_1t})\\
&+C\sigma\sqrt{d_1}((\log d_{\max})^{3/2}+\sqrt{t})\times\\
&\qquad\times\max_{i_1 i_3\dots i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|^{1/2}\\
&+C((\log d_{\max})^2+t)\max_{i_1,i_3,\dots,i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|\\
&+C((\log d_{\max})^2+t)\max_{i_2,i_3,\dots,i_p}\left|\sum_{i_1=1}^{d_1}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|.
\end{aligned}$$

with probability at least $1-L\exp(-t)$. The same bound holds for $\bigcup_{\mathbf{x}_q\in N^{(q)}_{L_q},2\leq q\leq p}\sum_{l=1}^{L}\mathbf{T}_l^\top\mathbf{U}_l$ as well. In the above, we have used the fact that $L:=\max L_q\leq\log d_{\max}+c_0$ for some constant $c>0$.

Integrating over t, one obtains, by (14), for $r\geq 1$,

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\mathbf{x}_q\in N^{(q)}_{L_q}:2\leq q\leq p}\left\|\sum_{l=1}^{L}\mathbf{T}_l(\mathbf{x}_2,\dots,\mathbf{x}_p)\right\|^r\big|\mathscr{E}\right)\\
&\leq C_p(\sigma^2d_{\max})^{r/2}\\
&\quad+C_p(\sigma^2d_1)^{\frac{r}{4}}(\log d_{\max})^{\frac{3r}{4}}\times\\
&\qquad\times\max_{i_1,i_3,\dots,i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|^{\frac{r}{4}}\\
&\quad+C_p(\log d_{\max})^r\times\\
&\quad\times\left\{\max_{i_1,i_3,\dots,i_p}\left|\sum_{i_2=1}^{d_2}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|\right.\\
&\qquad\left.+\max_{i_2,i_3,\dots,i_p}\left|\sum_{i_1=1}^{d_1}(E^2_{i_1\dots i_p}-\sigma^2_{i_1\dots i_p})\right|\right\}^{r/2}.
\end{aligned}$$

Summing over all terms in (13), we have then proved (12). Now taking expectation in (11) finishes the proof. □

B. Norm of Random Tensors

We are now in a position to prove Theorems 2.1 and 2.2. Without loss of generality, we take $\sigma=1$.

Proof of Theorem 2.1: We will use the definition of the geometric mean $\bar{d}_G=\left(\prod_{k=1}^{p}d_k\right)^{1/p}$. We begin with the upper bound.

Upper Bound: For any $t>d$, and $k\in[p]$, $\mathbb{E}\,\|*\|\,E^\alpha=\kappa<\infty$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{i_l\in[d_l],l\neq k}\sum_{i_k}\left(E^2_{i_1\dots i_p}-1\right)>t\right)\\
&\leq\prod_{l\neq k}d_l\cdot\frac{\mathbb{E}\left|\sum_{i_k}(E^2_{i_1\dots i_p}-1)\right|^{\alpha/2}}{t^{\alpha/2}}\\
&\leq\frac{C_p\kappa d_k^{\alpha/4}\prod_{l\neq k}d_l}{t^{\alpha/2}}
\end{aligned}$$

by Khintchine and Rosenthal inequalities respectively. This means

$$\mathbb{E}\left|\max_{i_l\in[d_l],l\neq k}\sum_{i_k=1}^{d_k}(E^2_{i_1\dots i_p}-1)\right|\leq C_pd_k^{1/2-2/\alpha}\left(\bar{d}_G\right)^{2p/\alpha}.$$

By Theorem 7.1 with $r=2$, and $\sigma^2_{i_1\dots i_p}=1$, we have

$$\begin{aligned}
&\mathbb{E}\,\|\mathscr{E}\|\\
&\leq C_p\sqrt{d_{\max}}+C_p(\log d_{\max})^{3/4}d_{\max}^{3/8-1/2\alpha}(\bar{d}_G)^{p/2\alpha}\\
&\quad+C_p(\log d_{\max})d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{p/\alpha}.
\end{aligned}$$

Notice that we can get a constant $C > 0$ such that

$$\begin{aligned}&\mathbb{P}\left(\max\left|E_{i_1\dots i_p}\right| > Cd_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{p/\alpha}\right)\\ \leq\ &\left(\prod_k d_k\right)\mathbb{E}\left\|*\right\|E^{\alpha}/\left(C^{\alpha}d_{\max}^{\alpha/4-1}\prod_k d_k\right)\\ =\ &d_{\max}^{1-\frac{\alpha}{4}}.\end{aligned}$$

It is well known that the function $f : \mathbb{R}^{d^p} \to \mathbb{R}$ given by $f(\mathrm{vec}(\mathscr{E})) = \|\mathscr{E}\|$ is convex and 1-Lipschitz. Now using Talagrand's concentration inequality for convex Lipschitz functions (see, e.g., Equation 1.4 of [43]) we obtain

$$\begin{aligned}&\mathbb{P}\Bigg(\left|\|\mathscr{E}\| - \mathbb{E}\|\mathscr{E}\|\right| > C_p\sqrt{d_{\max}}\\ &\qquad\qquad + C_p(\log d_{\max})^{3/2}d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{p/\alpha}\Bigg)\\ &\leq d_{\max}^{1-\frac{\alpha}{4}}\end{aligned}$$

and thus the upper bound now follows. In the last step, we use AM-GM inequality to remove the middle term. In general, this inequality is not sharp and leads to an extra $\sqrt{\log d}$ factor in the third term, but we keep this version for better presentation.

Now consider the lower bound.

Lower Bound: Suppose, without loss of generality, we have that $d_1 \geq d_2 \geq \dots d_p$. It is clear that

$$\|\mathscr{E}\|^2 \geq \max_{i_2,\dots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\dots i_p}^2.$$

Thus, for any constant $C > 0$,

$$\begin{aligned}&\mathbb{P}\left(\|\mathscr{E}\|^2 > d_1 + C^2\left(\bar{d}_G\right)^{\frac{2p}{\alpha}}d_{\max}^{\frac{1}{2}-\frac{2}{\alpha}}\text{ i.o.}\right)\\ \geq\ &\mathbb{P}\left(\max_{i_2,\dots,i_p}\sum_{i_1=1}^{d_1}E_{i_1\dots i_p}^2 - d_1 > C^4\left(\bar{d}_G\right)^{\frac{2p}{\alpha}}d_{\max}^{\frac{1}{2}-\frac{2}{\alpha}}\text{ i.o.}\right)\end{aligned}$$

Notice that $\sum_{i_1=1}^{d_1}(E_{i_1\dots i_p}^2 - 1)$ is a sum of independent mean zero random variables. Since $\mathbb{E}\left|E_{i_1\dots i_p}\right|^{\alpha/2} = \infty$, Corollary 2 of [41] along with Khintchine inequalities imply that for any finite d, the random variables

$$X_{i_2\dots i_p} = \sum_{i_1=1}^{d_1}(E_{i_1 i_2\dots i_p}^2 - 1)/\sqrt{d_1}$$

satisfy

$$\mathbb{E}|X_{i_2\dots i_p}|^{\alpha/2} \asymp \max\{1,\, d_1^{1-\alpha/4}(\mathbb{E}|E_{i_1 i_2\dots i_p}^2 - 1|)^{\alpha/2}\} = \infty.$$

For $k = 1, 2, \dots,$ let

$$B_k = \{i_2, \dots, i_p : 2^{k-1} < i_2 \leq 2^k, 1 \leq i_3, \dots, i_p \leq 2^k\}.$$

By Borel Cantelli theorem, it is enough to show that

$$\begin{aligned}&\sum_k\mathbb{P}\Bigg(\text{there exist } i_2,\dots,i_p \in B_k\\ &\qquad\qquad\text{s.t. }\left|X_{i_2\dots i_p}\right| \geq C\cdot 2^{2k(p-1)/\alpha}\Bigg) = \infty.\end{aligned}$$

In other words, we need

$$\sum_k\left[1 - \mathbb{P}\left(\left|X_{i_2\dots i_p}\right| < C\cdot 2^{2k(p-1)/\alpha}\right)^{2^{k(p-1)}/2}\right] = \infty.$$

Again since $\mathbb{E}\left|X_{i_2\dots i_p}\right|^{\alpha/2} = \infty$, we have $\mathbb{P}(|X_{i_1\dots i_p}| > t) \gtrsim t^{-\alpha/2}$ for large enough t, and hence

$$\sum_{k=1}^{\infty}2^{k(p-1)}\mathbb{P}\left(\left|X_{i_2\dots i_p}\right| \geq C\cdot 2^{2k(p-1)/\alpha}\right) = \infty. \tag{24}$$

We also use the well known implication

$$\sum_k\left[1 - (1 - a_k)^{b_k}\right] < \infty \implies \sum_k a_k b_k < \infty \tag{25}$$

for $a_k \in [0, 1],\ b_k \geq 0$.

Notice that if we define $\tilde{d}_G := (d_2 \dots d_p)^{1/(p-1)} = 2^k$, then equations (24) and (25) together imply that for any constant $C > 0$, tensors $\mathscr{E}$ of dimension $d_1 \times d_2 \times \dots \times d_p$ satisfy

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\mathscr{E}\|^2 - d_1}{\sqrt{d_1}} \geq C\cdot\tilde{d}_G^{\frac{2(p-1)}{\alpha}}\text{ for infinitely many }\tilde{d}_G\right)\\ &=1.\end{aligned}$$

Note that by assumption, the dimensions satisfy the relation $d_1 = d_{\max} \geq \tilde{d}_G = (d_2 \dots d_p)^{1/(p-1)}$. Thus,

$$\mathbb{P}\left(\|\mathscr{E}\|^2 \geq d_1 + C\sqrt{d_1}\left(\prod_{k=2}^p d_k\right)^{2/\alpha}\text{ i.o.}\right) = 1.$$

Since $d_1 = d_{\max}$, the proof is now completed for the case where, for example, $d_1, d_2 \to \infty$. If only $d_1 \to \infty$ and $d_2, \dots, d_p$ remain fixed, we can follow the same route, only the maximum over all possible choices of $i_2, \dots, i_p$ is only a finite maximum. In particular, we will have that

$$\mathbb{P}(\|\mathscr{E}\|^2 \geq Cd_1) = 1.$$

Since in this case $Cd_1 \geq Cd_1/2 + C\sqrt{d_1}\,(d_2 \dots d_p)^{2/\alpha}$, for sufficiently large d_1, the proof also follows in this case. □

The proof of Theorem 2.2 follows a similar strategy.

Proof of Theorem 2.2:

Upper Bound: Recall that $\mathbb{E}\left\|*\right\|E^{\alpha} = \kappa < \infty$. By Markov inequality,

$$\begin{aligned}&\mathbb{P}\left(\max_{i_l\in[d_l],l\neq k}\sum_{i_k=1}^{d_k}E_{i_1\dots i_p}^2 > t\right)\\ \leq\ &\left(\prod_{l\neq k}d_l\right)\cdot\mathbb{E}\left|\sum_{i_k=1}^{d_k}E_{i_1\dots i_p}^2\right|^{\alpha/2}/t^{\alpha/2}.\end{aligned}$$

By Khintchine's inequality, for independent Rademacher random variables R_{i_2},

$$\begin{aligned}
&\mathbb{E}\Big|\sum_{i_k=1}^{d_k} E_{i_1\dots i_p}^2\Big|^{\frac{\alpha}{2}}\\
&\le C\cdot\mathbb{E}\Big|\sum_{i_k=1}^{d_k} R_{i_k}E_{i_1\dots i_p}\Big|^{\alpha}\\
&\le C\cdot\max\left\{\left(\mathbb{E}\Big|\sum_{i_k=1}^{d_k} R_{i_k}E_{i_1\dots i_p}\Big|^2\right)^{\frac{\alpha}{2}}, d_k\mathbb{E}\big|E\big|^{\frac{\alpha}{2}}\right\}\\
&= Cd_k^{\frac{\alpha}{2}},
\end{aligned}$$

where the second step is by Rosenthal's inequality. Consequently

$$\mathbb{P}\left(\max_{i_l\in[d_l],l\neq q}\sum_{i_k=1}^{d_k}E_{i_1\dots i_p}^2>t\right)\le C\left(\prod_{l\neq k}d_l\right)d_k^{\frac{\alpha}{2}}/t^{\frac{\alpha}{2}},$$

and thus

$$\mathbb{E}\max_{i_l\in[d_l],l\neq q}\sum_{i_k=1}^{d_k}E_{i_1\dots i_p}^2\le C_p\left(\bar d_G\right)^{\frac{2p}{\alpha}}d_k^{1-2/\alpha}.$$

Similar to before, by Theorem 7.1 with $q=2$ and $\sigma_{i_1\dots i_p}^2=1$,

$$\mathbb{E}\,\|\mathscr{E}\|\le C\sqrt{d_{\max}}+C_p\left(\bar d_G\right)^{\frac{p}{\alpha}}d_k^{1/2-1/\alpha}(\log d)_{\max}^{3/2}.$$

Moreover, we can get a constant $C>0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\max\big|E_{i_1\dots i_p}\big|>C_p d_{\max}^{1/2-1/\alpha}\left(\bar d_G\right)^{\frac{p}{\alpha}}\right)\\
&\le\left(\prod_k d_k\right)\mathbb{E}\big|E\big|^{\alpha}/C^{\alpha}\left(C^{\alpha}d_{\max}^{\alpha/2-1}\prod_{k=1}^{p}d_k\right)=d_{\max}^{1-\frac{\alpha}{2}}.
\end{aligned}$$

Again, using Talagrand's concentration inequality for convex Lipschitz functions we obtain

$$\begin{aligned}
&\mathbb{P}\Bigg(\Big|\|\mathscr{E}\|-\mathbb{E}\|\mathscr{E}\|\Big|>C\sqrt{d_{\max}}\\
&\qquad\qquad+C_p(\log d_{\max})^{3/2}d_{\max}^{1/2-1/\alpha}\left(\bar d_G\right)^{\frac{p}{\alpha}}\Bigg)\\
&\le d_{\max}^{1-\frac{\alpha}{2}},
\end{aligned}$$

and the upper bound now follows.

Lower Bound: We will show that for any constant $C>0$,

$$\mathbb{P}\left(\|\mathscr{E}\|^2>C\left(\bar d_G\right)^{2p/\alpha}\text{ i.o.}\right)=1.$$

Clearly,

$$\|\mathscr{E}\|\ge\max_{i_1,\dots,i_p}\big|E_{i_1\dots i_p}\big|.$$

For $k=1,2,\dots,$ let

$$A_k=\{i_1,\dots,i_p:2^{k-1}<i_1\le 2^k,1\le i_2,\dots,i_p\le 2^k\}.$$

By Borel Cantelli theorem, it is enough to show that

$$\begin{aligned}
&\sum_k\mathbb{P}\Bigg(\text{there exist } i_1,\dots,i_p\in A_k\\
&\qquad\qquad\text{s.t. }\big|E_{i_1\dots i_p}\big|\ge C\cdot 2^{kp/\alpha}\Bigg)=\infty.
\end{aligned}$$

As before, we need

$$\sum_k\left[1-\mathbb{P}\left(\big|E_{i_1\dots i_p}\big|<C\cdot 2^{kp/\alpha}\right)^{2^{kp}/2}\right]=\infty.$$

Notice now that

$$\begin{aligned}
&\mathbb{E}\big|E_{i_1\dots i_p}\big|^{\alpha}\\
&\le\ C^{\alpha}+\sum_k\mathbb{E}(\big|E_{i_1\dots i_p}\big|^{\alpha}\mathbb{1}_{C\cdot 2^{kp/\alpha}\le\left|E_{i_1\dots i_p}\right|\le C\cdot 2^{(k+1)p/\alpha}}).
\end{aligned}$$

Thus,

$$\sum_{k=1}^{\infty}2^{kp}\mathbb{P}(C\cdot 2^{kp/\alpha}\le\big|E_{i_1\dots i_p}\big|\le C\cdot 2^{(k+1)p/\alpha})=\infty$$

since $\mathbb{E}\big|E_{i_1\dots i_p}\big|^{\alpha}=\infty$. Plugging in $\bar d_G=2^k$, the conclusion now follows from equation (25). Finally, we also have

$$\|\mathscr{E}\|^2\ge\max_{i_l\in[d_l],l\neq k}\sum_{i_1=1}^{d_k}E_{i_1\dots i_p}^2\ge Cd_k$$

almost surely for all $1\le k\le p$, and thus $\|\mathscr{E}\|\ge C\sqrt{d_{\max}}$ following the proof for the lower bound in Theorem 2.1. This completes the proof. □

C. Bounds for Tensor SVD

We now turn our attention to bounds for the tensor SVD and prove Theorems 2.3 and 2.4. The cases when $\alpha>4$ and $2<\alpha<4$ can be treated in an identical fashion and we shall focus on the case when $\alpha>4$ for brevity.

Proof: [Proof of Theorem 2.3]

Note that

$$\begin{aligned}
\widehat\lambda\ &:=\ \mathscr{X}\times_1\widehat{\mathbf{u}}_1^{\mathrm{SVD}}\cdots\times_p\widehat{\mathbf{u}}_p^{\mathrm{SVD}}\\
&\ge\ \mathscr{X}\times_1\mathbf{u}_1\cdots\times_p\mathbf{u}_p\\
&=\ \lambda+\mathscr{E}\times_1\mathbf{u}_1\cdots\times\mathbf{u}_p\\
&\ge\ \lambda-\sqrt{d_{\max}}
\end{aligned}\tag{26}$$

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

Write

$$\widehat{\mathbf{u}}_j^{\mathrm{SVD}}=\sqrt{1-\rho_j^2}\mathbf{u}_j+\rho_j\mathbf{v}_j$$

where $\|\mathbf{v}_j\|=1$ and $\mathbf{v}_j\perp\mathbf{u}_j$, for $1\le j\le p$. Let $\rho:=\max_j|\rho_j|$. Using the upper bounds from Theorem 2.1 for $k\ge 2$, we can derive that

$$\begin{aligned}
\widehat\lambda=&\mathscr{X}\times_1\widehat{\mathbf{u}}_1^{\mathrm{SVD}}\cdots\times_p\widehat{\mathbf{u}}_p^{\mathrm{SVD}}\\
=&\lambda\prod_{j=1}^{p}\sqrt{1-\rho_j^2}+\\
&+\sum_{A\subset[p],A\neq\emptyset}\left(\prod_{j\notin A}\sqrt{1-\rho_j^2}\right)\left(\prod_{k\in A}\rho_j\right)\times\\
&\qquad\left(\mathscr{E}\times_{j\notin A}\mathbf{u}_j\times_{k\in A}\mathbf{v}_k\right)
\end{aligned}$$

We bound the two terms in the sum separately. First,

$$\lambda\left(\prod_{j=1}^{p}\sqrt{1-\rho_j^2}\right)\le\lambda\sqrt{1-\rho^2}$$
$$\le\frac{\lambda}{2}(2-\rho^2)\le\lambda(1-\rho^2/2).$$

For the second term, we have two cases.

Case 1: $|A|=1$:

$$\begin{aligned}&\sum_{A\subset[p],|A|=1}\left(\prod_{j\notin A}\sqrt{1-\rho_j^2}\prod_{k\in A}\rho_j\right)\mathscr{E}\times_{j\notin A}\mathbf{u}_j\times_{k\in A}\mathbf{v}_k\\&\le\rho\sum_{A\subset[p],|A|=1}\|\mathscr{E}\times_{j\notin A}\mathbf{u}_j\|\\&\le C\rho\sqrt{d_{\max}}\end{aligned}$$

with probability at least $1-d_{\max}^{-1}$. Here we use Chebyshev inequality in the last line.

Case 2: $|A|\ge 2$:

$$\begin{aligned}&\sum_{\substack{A\subset[p]\\|A|\ge2}}\left(\prod_{j\notin A}\sqrt{1-\rho_j^2}\right)\left(\prod_{k\in A}\rho_j\right)\mathscr{E}\times_{j\notin A}\mathbf{u}_j\times_{k\in A}\mathbf{v}_k\\&\le\sum_{A\subset[p],|A|\ge2}\left(\prod_{j\in A}\sqrt{1-\rho_j^2}\right)\left(\prod_{j\notin A}\rho_j\right)\|*\|\,\mathscr{E}\times_{j\in A}\mathbf{u}_j\\&\le C_p\sum_{k=2}^{p}\rho^k\|\mathscr{E}\|\\&\le C_p\rho^2\left(\sqrt{d_{\max}}+d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{1/\alpha}(\log d_{\max})^{3/2}\right),\end{aligned}$$

with probability at least $1-d_{\max}^{1-\alpha/4}$, using the upper bounds from Theorem 2.1 in the last step. Combining all the terms, we have

$$\begin{aligned}\widehat{\lambda}\le{}&\lambda+C_p\sqrt{d_{\max}}\\&+\left[C_p\sqrt{d_{\max}}+C_pd_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{1/\alpha}(\log d_{\max})^{3/2}\right]\rho^2\\&-\lambda\rho^2/2\end{aligned}\tag{27}$$

with probability at least $1-d_{\max}^{-1}-d_{\max}^{1-\alpha/4}$. Note that $|\mathscr{E}\times_1\mathbf{u}_1\cdots\times_p\mathbf{u}_p|\le C\sqrt{d_{\max}}$ and moreover $\|\mathscr{E}\times_{k\neq j}\mathbf{u}_k\|\le C\sqrt{d_{\max}}$ with probability at least $1-d_{\max}^{-1}$, using Chebychev inequalities.

We can get a sufficiently large constant $C_p>0$, such that if

$$\lambda>\ C_p\left(\sqrt{d_{\max}}+d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{1/\alpha}(\log d_{\max})^{3/2}\right),$$

and $\rho^2>\sqrt{d_{\max}}/\lambda$, the last line of (27) is at most $\lambda-2\sqrt{d_{\max}}$, thus contradicting (26). We thus have

$$\rho^2\le\sqrt{d_{\max}}/\lambda.\tag{28}$$

It is also clear from (27) that $\widehat{\lambda}\le\lambda+C_p\sqrt{d_{\max}}$, which combined with (26) yields

$$|\widehat{\lambda}-\lambda|\le C_p\sqrt{d_{\max}}.$$

We will derive an improved upper bound on ρ by using the first order condition on $\widehat{\mathbf{u}}_j^{\rm SVD}$. In particular, $(\widehat{\mathbf{u}}_1^{\rm SVD},\ldots,\widehat{\mathbf{u}}_p^{\rm SVD})$ is a local minimum of the function

$$F(\gamma,\mathbf{a}_1,\ldots,\mathbf{a}_p)=\|\mathscr{X}-\gamma\mathbf{a}_1\circ\cdots\circ\mathbf{a}_p\|_{\rm HS}^2$$

for $\gamma\in\mathbb{R}$, $\mathbf{a}_j\in\mathbb{S}^{d-1}$. Setting the derivative of the Lagrangian to zero, we have

$$\mathscr{X}\times_{k\neq j}\widehat{\mathbf{u}}_k^{\rm SVD}=\widehat{\lambda}\widehat{\mathbf{u}}_j^{\rm SVD}\quad\text{for }1\le j\le p.$$

For $j=1$,

$$\begin{aligned}&\|\lambda(\widehat{\mathbf{u}}_1^{\rm SVD}-\mathbf{u}_1)\|\\&=\|(\lambda-\widehat{\lambda})\widehat{\mathbf{u}}_1^{\rm SVD}+(\widehat{\lambda}\widehat{\mathbf{u}}_1^{\rm SVD}-\lambda\mathbf{u}_1)\|\\&\le|\widehat{\lambda}-\lambda|+\|(\mathscr{T}+\mathscr{E})\times_{k\neq1}\widehat{\mathbf{u}}_k^{\rm SVD}-\lambda\mathbf{u}_1\|\\&\le C_pd_{\max}^{1/2}+\|*\|\,\lambda\big(\prod_{k\neq1}\sqrt{1-\rho_k^2}-1\big)\mathbf{u}_1+\|\mathscr{E}\times_{k\neq1}\widehat{\mathbf{u}}_k^{\rm SVD}\|.\end{aligned}\tag{29}$$

Since $\rho^2=\max_j\rho_j^2\le\sqrt{d_{\max}}/\lambda$ by (28), it is not hard to check that

$$\left|\prod_{k\neq1}\sqrt{1-\rho_k^2}-1\right|\le1-(1-\rho^2)^{(p-1)/2}\le C_p\sqrt{d_{\max}}/\lambda.$$

On the other hand, following the steps of (27), we have

$$\begin{aligned}&\|\mathscr{E}\times_{k\neq1}\widehat{\mathbf{u}}_k^{\rm SVD}\|\\&=\Bigg\|\sum_{A\subset([p]\setminus\{1\})}\left(\prod_{k\in A}\sqrt{1-\rho_j^2}\right)\left(\prod_{k\notin A\cup\{1\}}\rho_j\right)\times\\&\qquad\left(\mathscr{E}\times_{j\in A}\mathbf{u}_j\times_{k\notin A\cup\{1\}}\mathbf{v}_j\right)\Bigg\|\\&\le\sum_{A\subset([p]\setminus\{1\})}\left(\prod_{k\in A}\sqrt{1-\rho_j^2}\right)\left(\prod_{k\notin A\cup\{1\}}\rho_j\right)\|\mathscr{E}\times_{j\in A}\mathbf{u}_j\|\\&\le C_p\sqrt{d_{\max}}+C_p\rho\sqrt{d_{\max}}+C_p\sum_{k=2}^{p-1}\rho^k\|\mathscr{E}\|\\&\le C_p\sqrt{d_{\max}}\\&\quad+C_p\rho^2\left(\sqrt{d_{\max}}+d_{\max}^{1/4-1/\alpha}\left(\bar{d}_G\right)^{1/\alpha}(\log d_{\max})^{3/2}\right)\\&\le C_p\sqrt{d_{\max}}\\&\quad+C_p\cdot\frac{\sqrt{d_{\max}}}{\lambda}\cdot\left(\sqrt{d_{\max}}+d_{\max}^{\frac14-\frac1\alpha}\left(\bar{d}_G\right)^{\frac1\alpha}(\log d_{\max})^{\frac32}\right)\\&\le C_p\sqrt{d_{\max}},\end{aligned}$$

with probability at least $1-d_{\max}^{1-\alpha/4}$, once again using the upper bounds from Theorem 2.1. The last line uses the facts $\rho^2\le\sqrt{d_{\max}}/\lambda$ and our assumption that $\lambda>C\lambda_{\rm crit}(d_{\max},\bar{d}_G;\alpha)$ for a sufficiently large constant $C>0$.

Plugging the last two bounds into (29) above implies

$$\sin\angle(\widehat{\mathbf{u}}_1^{\rm SVD},\mathbf{u}_1)\le\sqrt{2}\|\widehat{\mathbf{u}}_1^{\rm SVD}-\mathbf{u}_1\|\le\frac{C_p\sqrt{d_{\max}}}{\lambda}.$$

The bounds for $j=2,\ldots,p$ follow by an analogous argument. ■

Proof: [Proof of Theorem 2.4] Consider for some constant $0 < C_0 < 1$, a set of vectors

$$\mathcal{S}(C_0) = \Big\{ \mathbf{x}_1, \ldots, \mathbf{x}_p : \mathbf{x}_j = \sqrt{1-\rho_j^2}\mathbf{u}_j + \rho_j \mathbf{v}_j, \\ \|\mathbf{v}_j\| = 1, \mathbf{v}_j \perp \mathbf{u}_j, |\rho_j| < C_0 \Big\}.$$

By assumption, $d_1 \geq \cdots \geq d_p$. There exists a β such that $\mathbb{E}|E_{i_1 \ldots i_p}|^\beta < \infty$ for all $i_1 \ldots i_p \in [d]$ and $\beta > 4$,

$$\left(\prod_{k=2}^{p-1} d_k\right)^{1/\beta} \leq \left(\prod_{k=2}^{p} d_k\right)^{1/\alpha}.$$

Following the steps of (27), for any $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathcal{S}(C_0)$,

$$\begin{aligned} &\mathscr{E} \times_1 \mathbf{x}_1 \cdots \times_p \mathbf{x}_p \\ &= \sum_{A \subset [p]} \left(\prod_{j \notin A} \sqrt{1-\rho_j^2}\right) \left(\prod_{k \in A} \rho_k\right) \mathscr{E} \times_{j \notin A} \mathbf{u}_j \times_{k \in A} \mathbf{v}_k \\ &\leq \sum_{k=0}^{p} \binom{p}{k} \max_{|A|=k} \prod_{k \in A} \rho_k \left\| \mathscr{E} \times_{j \notin A} \mathbf{u}_j \right\| \qquad (30) \end{aligned}$$

Case 1: $k \leq 1$: We bound the corresponding terms by

$$\left|\mathscr{E} \times_1 \mathbf{u}_1 \ldots \mathbf{u}_p\right| + p \max_{j \in [p]} \rho_j \|\mathscr{E} \times_{k \neq j} \mathbf{u}_k\| \leq C\sqrt{d_{\max}}$$

with probability at least $1 - d_{\max}^{-1}$, using Chebychev inequalities.

Case 2: $2 \leq k \leq p-1$: For any set $A \subset [p]$, note that $\mathscr{E} \times_{j \notin A} \mathbf{u}_j$ is a k order tensor with dimensions $\{d_k : k \in A\}$. We will then use Theorem 2.1 to write

$$\begin{aligned} &\left\| \mathscr{E} \times_{j \notin A} \mathbf{u}_j \right\| \\ &\leq C_p \sqrt{d_{\max}} + C_p \max_{k \in A} d_k^{1/4 - 1/\beta} \left(\prod_{k \in A} d_k\right)^{1/\beta} (\log d_{\max})^{3/2} \\ &\leq C_p \sqrt{d_1} + C_p d_1^{1/4} \left(\prod_{k=2}^{p-1} d_k\right)^{1/\beta} (\log d_1)^{3/2} \end{aligned}$$

with probability at least $1 - d_{\max}^{1-\beta/4}$, since $d_1 \geq d_2 \cdots \geq d_p$. Then,

$$\begin{aligned} &\sum_{k=2}^{p-1} \binom{p}{k} \max_{|A|=k} \left(\prod_{k \in A} \rho_k\right) \left\| \mathscr{E} \times_{j \notin A} \mathbf{u}_j \right\| \\ &\leq C_p \sqrt{d_1} + C_p \sum_{k=2}^{p-1} \binom{p}{k} C_0^k \sqrt{d_1} \\ &\quad + C_p d_1^{\frac{1}{4}} \left(\prod_{k=2}^{p-1} d_k\right)^{\frac{1}{\beta}} (\log d_1)^{\frac{3}{2}} \qquad (31) \end{aligned}$$

with probability at least $1 - d_{\max}^{1-\beta/4}$, as $d_1 \geq d_2 \geq \cdots \geq d_p$.

On the other hand, $\mathbb{E}\big|E\big|^\alpha = \infty$ for some value $\alpha < 4(p-1)$. Then by the lower bounds in Theorems 2.1, for any constant $C > 0$,

$$\|\mathscr{E}\| > C\sqrt{d_1} + C d_1^{1/4} \left(\prod_{k=2}^{p} d_k\right)^{\frac{1}{\alpha}}$$

almost surely. Plugging in the upper bounds into (30), we have

$$\begin{aligned} &\sup_{\mathbf{x}_j \in \mathcal{S}(C_0)} \mathscr{X} \times_1 \mathbf{x}_1 \cdots \times_p \mathbf{x}_p \\ &\leq \sup_{\mathbf{x}_j \in \mathcal{S}(C_0)} \mathscr{T} \times_1 \mathbf{x}_1 \cdots \times_p \mathbf{x}_p + \sup_{\mathbf{x}_j \in \mathcal{S}(C_0)} \mathscr{E} \times_1 \mathbf{x}_1 \cdots \times_p \mathbf{x}_p \\ &\leq \lambda + C_p \sqrt{d_1} + C_p d_1^{1/4} \left(\prod_{k=2}^{p-1} d_k\right)^{\frac{1}{\beta}} (\log d_1)^{3/2} + C_0^p \|\mathscr{E}\| \\ &\leq \lambda + C_p \sqrt{d_1} + C_p d_1^{1/4} \left(\prod_{k=2}^{p} d_k\right)^{\frac{1}{\alpha}} (\log d_1)^{3/2} + C_0^p \|\mathscr{E}\| \\ &\leq C_p \sqrt{d_1} + C_p d_1^{1/4} \left(\prod_{k=2}^{p} d_k\right)^{\frac{1}{\alpha}} (\log d_1)^{3/2} + C_0^p \|\mathscr{E}\| - 2\lambda \\ &\leq \|\mathscr{E}\| - 2\lambda, \qquad (32) \end{aligned}$$

since $\left(\prod_{k=2}^{p-1} d_k\right)^{1/\beta} \leq \left(\prod_{k=2}^{p} d_k\right)^{1/\alpha}$ and the signal value $\lambda < C_p\sqrt{d_1} + C_p d_1^{1/4} \left(\prod_{k=2}^{p} d_k\right)^{\frac{1}{\alpha}}$. Again,

$$\begin{aligned} &\mathscr{X} \times_1 \widehat{\mathbf{u}}_1^{\mathrm{SVD}} \times_2 \widehat{\mathbf{u}}_2^{\mathrm{SVD}} \times_3 \cdots \times_p \widehat{\mathbf{u}}_p^{\mathrm{SVD}} \\ &= \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{S}^{d-1}} \mathscr{X} \times_1 \mathbf{x}_1 \cdots \times_p \mathbf{x}_p \geq \|*\| \mathscr{E} - \lambda, \end{aligned}$$

with probability tending to one, as $d_{\max} \to \infty$, by Theorem 2.1. Thus when compared to (32) shows that the global maximizer $(\widehat{\mathbf{u}}_1^{\mathrm{SVD}}, \ldots, \widehat{\mathbf{u}}_p^{\mathrm{SVD}}) \notin \mathcal{S}(C_0)$. In particular, $\|\widehat{\mathbf{u}}_j^{\mathrm{SVD}} - \mathbf{u}_j\| > C_0$ with probability at least $1 - d_{\max}^{1-\beta/4}$, for any $C_0 < 1$.

The same proof goes through for the case $\alpha > 4(p-1)$ provided there is a small enough constant C such that $\lambda < C\sqrt{d_{\max}}$. Similarly, the case where $2 < \alpha < 4$ can be proved through the upper and lower bounds from Theorem 2.2. ■

D. Bounds for Spectral Initialization and Power Iteration

We now consider polynomial time computable estimates when $\alpha \geq 4$ by establishing bounds for spectral initialization and power iterations.

Proof: [Proof of Proposition 3.1.] Define $d_{-k} := \prod_{q \neq k} d_q$. Notice that $\mathsf{Mat}_k(\mathscr{E})$ is a $d_k \times d_{-k}$ matrix of i.i.d. random variables with mean 0 and variance 1. Also,

$$\begin{aligned} &\mathsf{Mat}_k(\mathscr{T})\mathsf{Mat}_k(\mathscr{E})^\top \\ &= \lambda \mathbf{u}_k (\mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_{k-1} \otimes \mathbf{u}_{k+1} \otimes \cdots \otimes \mathbf{u}_p)^\top \mathsf{Mat}_k(\mathscr{E})^\top \\ &= \lambda \mathbf{u}_k (\mathbf{E}')^\top, \end{aligned}$$

where $\mathbf{E}'$ is a d_k length vector with independent random variables $\mathbb{E}\mathbf{E}_i' = 0$, $\mathbb{E}(\mathbf{E}_i')^2 = 1$ and $\mathbb{E}(\mathbf{E}_i')^4 = \kappa < \infty$. Then

$$\begin{aligned} &\mathbb{P}\left(\left\|\mathbf{u}_k(\mathbf{E}')^\top\right\| > 2C\sqrt{d_k}\right) \\ &= \mathbb{P}\left(\left(\sum E_i'^2 - 1\right)^2 > 4C d_k^2\right) \\ &\leq C d_k \mathbb{E}((E_i')^4)/d_k^2 \leq d_k^{-1}. \end{aligned}$$

By Bai-Yin's law, $\lambda_{\max}(\mathsf{Mat}_k(\mathscr{E})) = \left(\prod_{q\neq k} d_q\right)^{1/2} + \sqrt{d_k} + o(\sqrt{d_k})$ when $d_k/d_{-k} \to c \in [0,1]$. See, e.g., Theorem 2 of [44] and Theorem 5.31 of [45]. Note that the aspect ratio condition on d_k/d_{-k} is satisfied whenever $d_k < d_{\max}$ and also if $d_k = d_{\max}$ but $d_k/d_{-k} \to c \in [0,1]$.

If $d_k = d_{\max}$ and $d_k/d_{-k} \to c > 1$, by Theorem 2 of [37], we have that

$$\|\mathrm{Mat}_k(\mathscr{E})\| \leq C\sqrt{d_{\max}}$$

for large enough $d_{\max}$. Combining the two cases, we thus have, regardless of the aspect ratio d_k/d_{-k}, that

$$\begin{aligned}
&\left\|\mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{E})^\top - d_{-k}I_{d_k}\right\| \\
&= \lambda_{\max}(\mathsf{Mat}_k(\mathscr{E}))^2 - \prod_{q\neq k} d_q \\
&\leq C\max\left\{\left(\prod_{q=1}^{p} d_q\right)^{1/2}, d_k\right\}
\end{aligned}$$

almost surely as $d_{\max} \to \infty$.

Now using Davis-Kahan theorem,

$$\begin{aligned}
\sin\angle\left(\widehat{\mathbf{u}}_k^{\mathsf{Mat}}, \mathbf{u}_k\right) \leq& \frac{2\left\|\mathsf{Mat}_k(\mathscr{E})\mathsf{Mat}_k(\mathscr{E})^\top - d_{-k}I_{d_k}\right\|}{\lambda^2} \\
&+ \frac{4\lambda\left\|\mathsf{Mat}_k(\mathscr{T})\mathsf{Mat}_k(\mathscr{E})^\top\right\|}{\lambda^2} \\
\leq& \frac{Cd_k + \left(\prod_{q=1}^{p} d_q\right)^{1/2} + C\lambda\sqrt{d_k}}{\lambda^2}.
\end{aligned}$$

with probability at least $1 - d_k^{-1}$. The proof for other modes follows similarly. ■

Proof: [Proof of Proposition 3.2] The proof is by induction on t. The basis step holds by some nontrivial initialization, for example through the matricization estimator of Proposition 3.1. We now assume that the induction hypothesis holds for some $t > 0$ and prove the induction step for $t+1$.

As before, we write

$$\mathbf{x}_j^{[t]} = \sqrt{1-\rho_j^2}\mathbf{u}_j + \rho_j\mathbf{v}_j$$

where $\mathbf{v}_j$ is a unit length vector perpendicular to $\mathbf{u}_j$.

Then

$$\begin{aligned}
&\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]} \\
&= \lambda\left(\prod_{j\neq k}\sqrt{1-\rho_j^2}\right)\mathbf{u}_k \\
&\quad + \sum_{A\subset([p]\setminus\{k\})}\left(\prod_{j\in A}\sqrt{1-\rho_j^2}\prod_{l\notin A\cup\{k\}}\rho_j\right)\times \\
&\qquad \times\left(\mathscr{E} \times_{j\in A} \mathbf{u}_j \times_{l\notin A\cup\{k\}} \mathbf{v}_l\right). \qquad (33)
\end{aligned}$$

Notice that the entries of $\mathbf{E}' = \mathscr{E} \times_{j\neq k} \mathbf{u}_j$ are i.i.d. copies of a random variable E'' with $\mathbb{E}(E'') = 0$, $\mathrm{Var}(E'') = 1$ and $\mathbb{E}\left|E''\right|^4 = \kappa < \infty$. By Chebyshev's inequality, for any $1 \leq k \leq p$,

$$\begin{aligned}
&\mathbb{P}\left(\|\mathscr{E} \times_{j\neq k} \mathbf{u}_j\| > C\sqrt{d_k}\right) \\
&= \mathbb{P}\left(\left(\sum E_i'^2 - 1\right)^2 > 4Cd_k^2\right) \\
&\leq d_k\,\mathbb{E}((E_i')^4)/Cd_k^2 \leq d_k^{-1}.
\end{aligned}$$

Notice also that

$$\begin{aligned}
&\sum_{\substack{A\subset([p]\setminus\{k\}),\\ |A|\leq p-2}}\left(\prod_{j\in A}\sqrt{1-\rho_j^2}\right)\left(\prod_{l\notin A\cup\{k\}}\rho_j\right)\times \\
&\qquad\left(\mathscr{E} \times_{j\in A} \mathbf{u}_j \times_{l\notin A\cup\{k\}} \mathbf{v}_l\right) \\
&\leq \sum_{l=2}^{p-1}\binom{p}{l}(\rho^{[t]})^l\|\mathscr{E}\| \leq C_p(\rho^{[t]})^2\|\mathscr{E}\|.
\end{aligned}$$

The last two inequalities together imply that

$$\begin{aligned}
&\left\|\mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[t]}\right\| \\
&=\left\|\sum_{A\subset([p]\setminus\{k\})}\left(\prod_{j\in A}\sqrt{1-\rho_j^2}\right)\left(\prod_{j\notin A\cup\{k\}}\rho_j\right)\times\right. \\
&\qquad\left.\left(\mathscr{E} \times_{j\in A} \mathbf{u}_j \times_{j\notin A\cup\{k\}} \mathbf{v}_j\right)\right\| \\
&\leq C\sqrt{d_k} + C_p(\rho^{[t]})^2\|\mathscr{E}\|. \qquad (34)
\end{aligned}$$

By the nontrivial initialization and the induction hypothesis, we have a constant $\rho^* < 1$ such that $\rho^{[t]} \leq \rho_* < 1$. We then have

$$\begin{aligned}
&\sin\angle(\mathbf{x}_k^{[t]}, \mathbf{u}_k) \\
&= \sup_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{u}_k}\frac{\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]} \times_k \mathbf{w}}{\|\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]}\|} \\
&= \sup_{\|\mathbf{w}\|=1,\mathbf{w}\perp\mathbf{u}_k}\frac{\mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[t]} \times_k \mathbf{w}}{\|\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[t]}\|} \\
&\leq\frac{\left\|\mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[t]}\right\|}{\lambda\left(\prod_{j\neq k}\sqrt{1-\rho_j^2}\right) - \|*\|\,\mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[t]}} \\
&\leq\frac{C\sqrt{d_k} + C_p(\rho^{[t]})^2\|\mathscr{E}\|}{\lambda(1-\rho_*^2))^{(p-1)/2} - C\sqrt{d_{\max}} - C_p(\rho^{[t]})^2\|\mathscr{E}\|} \\
&\leq C\frac{\sqrt{d_k}}{\lambda} + C_p(\rho^{[t]})^2\frac{\|\mathscr{E}\|}{\lambda}
\end{aligned}$$

with probability at least $1 - d_k^{1-\alpha/4}$.

We use (33) and (34) for the first and second inequalities respectively. The last line follows if $\lambda > C\|\mathscr{E}\|$ for a sufficiently large constant $C > 0$. Since we have $\lambda > Cd_{\max}^{1/2} + C\left(\prod d_k\right)^{1/4}$ and $\alpha > 4$, this condition is satisfied with probability at least $1 - pd_{\min}^{1-\alpha/4}$, using the upper bounds from Theorem 2.1. ■

E. Bounds for Robust Tensor SVD

Proof: [Proof of Theorem 4.1.] Let us fix $k = 1$ as the other modes follow by symmetry.

We will denote the partition of $[d_{-1}]$ into n groups as $I := (I_1, \dots, I_n)$. Let us also define

$$\sigma^2 = \left\| \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}(\mathbf{S}_j^2 | I) \right\|.$$

Then conditional on I, we apply Theorem 3.2 of [34] with

$$\theta = \frac{\sqrt{2\log(2d_1/\delta)/n}}{\sigma}$$

to obtain

$$\mathbb{P}\left(\left\| \widehat{\mathbf{V}}_1 - \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left(\mathbf{S}_j | I\right) \right\| > \sigma\sqrt{\frac{2\log(2d_1/\delta)}{n}} \middle| I \right) \leq \delta. \tag{35}$$

We need to calculate $\mathbb{E}\left(\mathbf{S}_j \middle| I\right)$ and σ^2. We have the following lemma, which is proved in section .

Lemma 1: $\mathbb{E}(\mathbf{S}_j|I) = \left(\sum_{i_{-1} \in I_j} \mathbf{w}_{i_{-1}}^2 \right) \mathbf{V}_1$ and

$$\sigma^2 = \left\| \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}(\mathbf{S}_j^2|I) \right\| \leq \lambda^4 \left(\frac{1}{n}\sum_{j=1}^{n} \left(\sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2 \right)^2 \right) + \frac{\lambda^2 d_1}{n} \sum_{j=1}^{n} \sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2 + \frac{d_1}{n}\sum_{j=1}^{n} |I_j|. \tag{36}$$

To complete the proof, we now use the multinomial sample splitting scheme to get high probability bounds on the above quantities. We write $G_j = \sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2$. By our sampling scheme, $\mathbb{1}(i_{-1} \in I_j) \sim$ Bernoulli$(\frac{1}{n})$ so that $\mathbb{E}G_j = \frac{1}{n}$. Then, the scaled Chernoff bound (see, e.g., Theorems 1 and 2 and the subsequent remark of [40]) along with the definition of μ_1 yields,

$$\begin{aligned} &\mathbb{P}\left(\max_{1\leq j\leq n} G_j > \frac{2}{n} \right) \\ &\leq n\mathbb{P}\left(\frac{|G_1 - \mathbb{E}G_1|}{\max_{i_{-1}\in I_1} \|\mathbf{w}_{i_{-1}}\|^2} > \frac{\mathbb{E}G_1}{\max_{i_{-1}\in I_1} \|\mathbf{w}_{i_{-1}}\|^2} \right) \\ &\leq n\exp\left(-\frac{1/n}{\mu_1^2} \right). \end{aligned} \tag{37}$$

By the sample partition scheme

$$\sum_{j=1}^{n} |I_j| = \prod_{k=2}^{p} d_k \quad \text{and} \quad \sum_{j=1}^{n}\sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2 = \|\mathbf{w}\|^2 = 1.$$

Using (36) and (37) we then have

$$\sigma^2 \leq \frac{2\lambda^4}{n^2} + \frac{\lambda^2 d_1}{n} + \frac{1}{n}\cdot\prod_{k=1}^{p} d_k \tag{38}$$

with probability at least $1 - 2n\exp\left(-\frac{1}{2n\mu_1^2}\right)$. Notice also that the signal matrix is

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}(\mathbf{S}_j|I) &= \frac{1}{n}\left(\sum_{j}\sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2 \right)\mathbf{V}_1 \\ &= \frac{\lambda^2}{n}(\mathbf{u}_k\mathbf{u}_k^\top - \mathrm{diag}(\mathbf{u}_k\mathbf{u}_k^\top)). \end{aligned}$$

Now applying (35) with $\delta = 1/d_{\max}$, together with the noise bound from (38) and using Davis-Kahan theorem, we have

$$\begin{aligned} &\sin\angle(\widehat{\mathbf{v}}_k,\, \mathbf{u}_k) \\ &\leq 2\|\mathbf{u}_k\|_{\ell_\infty}^2 + \frac{2\sigma}{\lambda^2/n}\sqrt{\frac{2\log d_{\max}}{n}} \\ &\leq 2\mu_2^2 + 2\left(2 + \frac{(\lambda\sqrt{d_1} + (\prod_k d_k)^{\frac{1}{2}})\sqrt{n}}{\lambda^2} \right)\sqrt{\frac{2\log d_{\max}}{n}} \\ &\leq 2\mu_2^2 + 4\sqrt{\frac{2\log d_{\max}}{n}} + \frac{(\lambda\sqrt{d_1} + (\prod_k d_k)^{\frac{1}{2}})\sqrt{8\log d_{\max}}}{\lambda^2} \end{aligned}$$

with probability at least $1 - 2n\exp\left(-\frac{1}{2n\mu_1^2}\right) - \frac{1}{d_{\max}}$. ■

Proof: [Proof of Proposition 4.2]

Since $\mathbf{x}_k^{[0]}$ are unit vectors that are independent of $\mathscr{E}$, $\mathbf{E} = \mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[0]}$ is a $d_k \times 1$ vector whose entries are independent random variables with $\mathbb{E}E_i = 0$ and $\mathbb{E}E_i^2 = 1$. Moreover, $\mathbb{E}|E_i|^\alpha < \infty$ by Rosenthal inequality. Thus

$$\begin{aligned} &\mathbb{P}\left(\sup_{\mathbf{v}:\|\mathbf{v}=1\|} \mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[1]} \times_k \mathbf{v} > C\sqrt{d_k}/\delta^{1/\alpha} \right) \\ &\leq \mathbb{P}\left(\|\mathbf{E}\| > C\sqrt{d_k}/\delta^{1/\alpha} \right) \\ &\leq \frac{(\sum_i \mathbb{E}E_i^2)^{\alpha/2}}{Cd_k^{\alpha/2}/\delta} = \delta, \end{aligned}$$

where we use Rosenthal inequalities in the last step Therefore

$$\begin{aligned} \sin\angle\left(\mathbf{x}_k^{[1]}, \mathbf{u}_k\right) &= \sup_{\mathbf{v}:\|\mathbf{v}\|=1,\, \mathbf{v}\perp\mathbf{u}_k} |\langle \mathbf{x}_k^{[1]}, \mathbf{v}\rangle| \\ &\leq \sup_{\mathbf{v}:\|\mathbf{v}\|=1,\, \mathbf{v}\perp\mathbf{u}_k} \frac{\mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[0]} \times_k \mathbf{v}}{\|\mathscr{X} \times_{j\neq k} \mathbf{x}_j^{[0]}\|} \\ &\leq \frac{\left\| \mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[0]} \right\|}{\lambda\prod_{j\neq k} |\langle \mathbf{x}_j^{[0]}, \mathbf{u}_j\rangle| - \left\| \mathscr{E} \times_{j\neq k} \mathbf{x}_j^{[0]} \times_k \mathbf{u}_k \right\|} \\ &\leq \frac{C\sqrt{d_k}/\delta^{1/\alpha}}{\lambda(1-\eta^2)^{(p-1)/2} - \sqrt{d}/\delta^{1/\alpha}} \\ &\leq \frac{C\sqrt{d_k}/\delta^{1/\alpha}}{\lambda(1-\eta^2)^{(p-1)/2}}. \end{aligned}$$

with probability at least $1 - \delta$, provided $\lambda(1-\eta^2)^{(p-1)/2} \geq 2\sqrt{d_{\max}}/\delta^{1/\alpha}$. ■

Proof: [Proof of Theorem 4.3] To obtain estimates for the k-th mode vector $\mathbf{u}_k$, we split the tensor into two halves along the k-th mode, as described in Section IV. From the first half, we obtain initial estimators $\mathbf{v}_q$ for all $q \neq k$.

Note that the vector $\mathbf{u}_{k,J_2}$ is the k-th mode of $\mathscr{X}_2$. By the scaled Chernoff bounds (see, e.g., Theorems 1, 2 and subsequent remark of [40]),

$$\begin{aligned}&\mathbb{P}\left(\|\mathbf{u}_{k,J_2}\| < \frac{1}{2}\right)\\ &= \mathbb{P}\left(\left|\sum_{i=1}^{d_k} \left((\mathbf{u}_k)_i\right)^2 \mathbb{1}(B_i = 0) - \frac{1}{2}\right| > \frac{1}{4}\right)\\ &\leq \exp\left(-\frac{1/16}{\max\limits_i \left((\mathbf{u}_k)_i\right)^2}\right) \leq C d_k^{-1}\end{aligned}$$

so that $\|\mathbf{u}_{k,J_2}\| \geq 0.5$ with high probability. By Theorem 4.1 we have initializations $\mathbf{v}_q$ independent of $\mathscr{X}_2$, satisfying (10) for some constant $\eta < 1$.

We immediately have from proposition 4.2 that

$$\mathbb{P}\left(\sin\angle(\widehat{\mathbf{u}}_{k,J_2}, \mathbf{u}_{k,J_2}/\|\mathbf{u}_{k,J_2}\|) \leq \frac{C\sqrt{d_k}}{\lambda t}\right) \geq 1 - t^{\alpha}.$$

Finally, initializing with $\mathscr{X}_2$ and using $\mathscr{X}_1$ for optimal estimation, we also have $\widehat{\mathbf{u}}_{k,J_1}$ that is a rate optimal estimator of $\mathbf{u}_{k,J_1}$. Concatenating the two estimators to get $\widehat{\mathbf{u}}_k = \left(\widehat{\mathbf{u}}_{k,J_1}^{\top}\, \widehat{\mathbf{u}}_{k,J_2}^{\top}\right)^{\top}$, we get

$$\mathbb{P}\left(\sin\angle(\widehat{\mathbf{u}}_k, \mathbf{u}_k) \leq \frac{C\sqrt{d_k}}{\lambda t}\right) \geq 1 - 2t^{\alpha}.$$

We now show that it is possible to have improved estimators by splitting the tensor only once (say along the pth mode), if the signal strength is higher. Suppose $d_1 \geq d_2 \geq \cdots \geq d_p$ and $\lambda \geq C d_1^{1/2} d_p^{1/\alpha}$. Note that this is satisfied by the initialization condition $\lambda \geq C\left(\bar{d}_G\right)^{p/4} (\log \bar{d}_G)^{1/4}$, for example when $d_1 \leq C d_p$ and $p \geq 4$.

We consider the first mode $\mathbf{u}_1$. Since the unit initialization vectors $\widehat{\mathbf{v}}_k$ are independent of $\mathscr{X}_2$ for $k = 2, \ldots, p-1$ the matrix $\mathbf{E}_2 = \mathscr{E}_2 \times_2 \widehat{\mathbf{v}}_k \cdots \times_{p-1} \widehat{\mathbf{v}}_{p-1}$ satisfies $\mathbf{E}_{ij}$ are independent, $\mathbb{E}\mathbf{E}_{ij} = 0$, $\mathbb{E}\mathbf{E}_{ij}^2 = 1$. By Rosenthal inequalities, $\mathbb{E}|E_{ij}|^{\alpha} < \infty$.

Moreover $\mathbf{w}_2 = \mathbf{E}\mathbf{u}_{p,J_2}/\|\mathbf{u}_{p,J_2}\|$ again has independent entries with the same properties.

Suppose $d_1 \geq d_p$ without loss of generality. By Theorem 7.1 and Talagrand's concentration inequality, we have

$$\mathbb{P}\left(\|\mathbf{E}\| > C d_1^{\frac{1}{2} + \frac{1}{\alpha \mathbb{1}(\alpha<4)}}/t\right) \leq t^{\alpha}.$$

Under this event,

$$\begin{aligned}&\sup_{\mathbf{v}:\|\mathbf{v}\|=1,\, \mathbf{v}\perp\mathbf{u}^{(1)}} \mathscr{E}_2 \times_1 \mathbf{v} \times_2 \widehat{\mathbf{v}}_2 \cdots \times_{p-1} \widehat{\mathbf{v}}_{p-1} \times_p \widehat{\mathbf{u}}_{p,J_2}\\ &\leq \|\mathbf{E}(\mathbf{u}_{p,J_2}/\|\mathbf{u}_{p,J_2}\|)\| + \|\mathbf{E}\|\|\widehat{\mathbf{u}}_{p,J_2} - (\mathbf{u}_{p,J_2}/\|\mathbf{u}_{p,J_2}\|)\|\\ &\leq \|\mathbf{w}_2\| + \|\mathbf{E}\| \cdot \frac{C\sqrt{d_p}}{\lambda}\\ &\leq C\sqrt{d_1/t^2} + C \cdot \frac{d_1^{\frac{1}{2}} d_p^{\frac{1}{\alpha\mathbb{1}(\alpha<4)}}}{t} \cdot \frac{\sqrt{d_p}}{\lambda} \leq C\sqrt{d_1}/t.\end{aligned}$$

with probability at least $1 - t^{\alpha}$. The second last inequality uses the upper bounds on $\|\mathbf{w}\|$ and $\|\mathbf{E}\|$. The last inequality now follows since $\lambda > C d_1^{\frac{1}{2}} d_p^{\frac{1}{\alpha\mathbb{1}(\alpha<4)}}$. Hence for any $\delta > 0$,

$$\begin{aligned}&\sup_{\mathbf{v}:\|\mathbf{v}\|=1,\, \mathbf{v}\perp\mathbf{u}^{(1)}} \frac{\mathscr{E}_2 \times_1 \mathbf{v} \times_2 \widehat{\mathbf{v}}_2 \cdots \times_{p-1} \widehat{\mathbf{v}}_{p-1} \times_p \widehat{\mathbf{u}}_{p,J_2}}{\|\mathscr{X}_2 \times_2 \widehat{\mathbf{v}}_2 \cdots \times_{p-1} \widehat{\mathbf{v}}_{p-1} \times_p \widehat{\mathbf{u}}_{p,J_2}\|}\\ &\leq \frac{C\sqrt{d_1}/t}{\lambda\|\mathbf{u}_{p,J_2}\|(1-\eta^2)^{(p-1)/2} - C\sqrt{d_1}/t} \leq \frac{C\sqrt{d_1}}{\lambda t}\end{aligned}$$

with probability at least $1 - t^{\alpha}$. The proof for the rest of the modes follows similarly. Finally, initializing with $\mathscr{X}_2$ and using $\mathscr{X}_1$ for optimal estimation, we also have $\widehat{\mathbf{u}}_{p,J_1}$ that is a rate optimal estimator of $\mathbf{u}_{p,J_1}$. This finishes the proof. ■

Appendix

Proof: [Proof of Lemma 1] We write $\mathbf{w} = (\mathbf{u}_2 \otimes \mathbf{u}_3 \ldots \otimes \mathbf{u}_p)$. By definition

$$\mathbf{S}_j = \sum_{i_{-1}\in I_j} \left(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top} - \operatorname{diag}(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top})\right).$$

Notice that

$$\mathbb{E}(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top}) = \lambda^2 \mathbf{w}_{i_{-1}}^2 \mathbf{u}_1 \circ \mathbf{u}_1 + I,$$

which implies

$$\begin{aligned}\mathbb{E}(\mathbf{S}_j|I) &= \sum_{i_{-1}\in I_j} \lambda^2 \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1 \circ \mathbf{u}_1 - \operatorname{diag}(\mathbf{u}_1 \circ \mathbf{u}_1))\\ &= \left(\sum_{i_{-1}\in I_j} \mathbf{w}_{i_{-1}}^2\right) \mathbf{V}_1.\end{aligned}$$

Next, for any $i_{-1} \in I_j$ and $s, t \in [d_1]$, $s \neq t$

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top} - \operatorname{diag}(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top})\right)_{st}^2\right]\\ =&\mathbb{E}\left(\sum_{l\neq s,\, t} (\mathbf{X}_{i_{-1}})_s (\mathbf{X}_{i_{-1}})_t (\mathbf{X}_{i_{-1}})_l^2\right)\\ =&\mathbb{E}(\mathbf{X}_{i_{-1}})_s \mathbb{E}(\mathbf{X}_{i_{-1}})_t \sum_{l\neq s,\, t} \mathbb{E}(\mathbf{X}_{i_{-1}})_l^2\\ =&\lambda^2 \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1)_s (\mathbf{u}_1)_t \sum_{l\neq s,\, t} (\lambda^2 \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1)_l^2 + 1)\\ =&\lambda^4 \mathbf{w}_{i_{-1}}^4 (\mathbf{u}_1)_s (\mathbf{u}_1)_t (1 - (\mathbf{u}_1)_s^2 - (\mathbf{u}_1)_t^2)\\ &+ \lambda^2 (d_1 - 2) \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1)_s (\mathbf{u}_1)_t.\end{aligned}$$

On the other hand,

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top} - \operatorname{diag}(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^{\top})\right)_{ss}^2\right]\\ =&\mathbb{E}\left(\sum_{l\neq s} (\mathbf{X}_{i_{-1}})_s^2 (\mathbf{X}_{i_{-1}})_l^2\right)\\ =&(\lambda^2 \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1)_s^2 + 1) \sum_{l\neq s} (\lambda^2 \mathbf{w}_{i_{-1}}^2 (\mathbf{u}_1)_l^2 + 1)\\ =&\lambda^4 \mathbf{w}_{i_{-1}}^4 (\mathbf{u}_1)_s^2 (1 - (\mathbf{u}_1)_s^2)\\ &+ \lambda^2 \mathbf{w}_{i_{-1}}^2 (1 + (d_1 - 2)(\mathbf{u}_1)_s^2) + d_1 - 1.\end{aligned}$$

Collecting all the terms,

$$\begin{aligned}&\mathbb{E}[\Big(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^\top - \operatorname{diag}(\mathbf{X}_{i_{-1}}\mathbf{X}_{i_{-1}}^\top)\Big)^2] \\ &= \mathbf{w}_{i_{-1}}^4\mathbf{V}_1^2 + \lambda^2\mathbf{w}_{i_{-1}}^2(d_1-2)\mathbf{u}_1\mathbf{u}_1^\top \\ &\quad + [(d_1-1)+\lambda^2\mathbf{w}_{i_{-1}}^2]\mathbb{I}_{d_1}.\end{aligned}$$

Similarly for $i_{-1,1} \neq i_{-1,2} \in I_j$, and indices $s, t \in [d_1]$,

$$\begin{aligned}&\mathbb{E}\Big[\Big(\mathbf{X}_{i_{-1,1}}\mathbf{X}_{i_{-1,1}}^\top - \operatorname{diag}(\mathbf{X}_{i_{-1,1}}\mathbf{X}_{i_{-1,1}}^\top)\Big) \\ &\quad \Big(\mathbf{X}_{i_{-1,2}}\mathbf{X}_{i_{-1,2}}^\top - \operatorname{diag}(\mathbf{X}_{i_{-1,2}}\mathbf{X}_{i_{-1,2}}^\top)\Big)\Big]_{st} \\ &= \lambda^4\mathbf{w}_{i_{-1,1}}^2\mathbf{w}_{i_{-1,2}}^2(\mathbf{u}_1)_s(\mathbf{u}_1)_t(1-(\mathbf{u}_1)_s^2 \\ &\quad -(\mathbf{u}_1)_t^2\mathbb{1}(s\neq t))\end{aligned}$$

meaning

$$\begin{aligned}&\mathbb{E}\Big[\Big(\mathbf{X}_{i_{-1,1}}\mathbf{X}_{i_{-1,1}}^\top - \operatorname{diag}(\mathbf{X}_{i_{-1,1}}\mathbf{X}_{i_{-1,1}}^\top)\Big) \\ &\quad \Big(\mathbf{X}_{i_{-1,2}}\mathbf{X}_{i_{-1,2}}^\top - \operatorname{diag}(\mathbf{X}_{i_{-1,2}}\mathbf{X}_{i_{-1,2}}^\top)\Big)\Big] \\ &= \mathbf{w}_{i_{-1,1}}^2\mathbf{w}_{i_{-1,2}}^2\mathbf{V}_1^2.\end{aligned}$$

Adding the terms above,

$$\begin{aligned}&\mathbb{E}(\mathbf{S}_j^2|I) \\ &= \Bigg(\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2\Bigg)^2\mathbf{V}_1^2 + \lambda^2(d_1-2)\mathbf{u}_1\mathbf{u}_1^\top\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2 \\ &\quad + \sum_{i_{-1}\in I_j}[(d_1-1)+\lambda^2\mathbf{w}_{i_{-1}}^2]\mathbb{I}_{d_1} \\ &= \Bigg(\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2\Bigg)^2\mathbf{V}_1^2 + (d_1-1)|I_j|\mathbb{I}_{d_1} \\ &\quad + \lambda^2\left[(d_1-2)\mathbf{u}_1\mathbf{u}_1^\top + \mathbb{I}_{d_1}\right]\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2.\end{aligned}$$

Consequently, conditional on I,

$$\begin{aligned}\sigma^2 &= \left\|\frac{1}{n}\sum_{j=1}^n\mathbb{E}(\mathbf{S}_j^2|I)\right\| \\ &= \frac{1}{n}\Bigg\|\sum_{j=1}^n\Bigg[\Bigg(\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2\Bigg)^2\mathbf{V}_1^2 \\ &\qquad + \lambda^2\left[(d_1-2)\mathbf{u}_1\mathbf{u}_1^\top+\mathbb{I}_{d_1}\right]\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2 \\ &\qquad + (d_1-1)|I_j|\mathbb{I}_d\Bigg]\Bigg\| \\ &\le \lambda^4\Bigg(\frac{1}{n}\sum_{j=1}^n\Bigg(\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2\Bigg)^2\Bigg) \\ &\quad + \frac{\lambda^2 d_1}{n}\sum_{j=1}^n\sum_{i_{-1}\in I_j}\mathbf{w}_{i_{-1}}^2 + \frac{d_1}{n}\sum_{j=1}^n|I_j|.\end{aligned}$$

This finishes the proof. ■

Lemma 2: There exist a constant C_a such that $\sum_{l=0}^{L}2^{al}$ $(1+L-l) \le C_a \cdot 2^{aL}$ for $a > 0$.

Proof: Note that the function $f(l) = 2^{al}(1+L-l)$ is increasing when $0 \le l \le L$. Thus we can bound

$$\begin{aligned}&\sum_{l=0}^{L}2^{al}(1+L-l) \\ &\le \int_0^L 2^{al}(1+L-l)dl = \int_1^{\exp(L)} t^{a\log 2-1}(1+L-\log t)dt \\ &= (1+L)[\exp(aL\log 2)-1] - \frac{1}{a\log 2}\int_1^{e^{aL\log 2}}\log(u)du \\ &= (1+L)[\exp(aL\log 2)-1] \\ &\quad - \frac{1}{a\log 2}\left[\exp(aL\log 2)(aL\log 2) - \exp(aL\log 2)+1\right] \\ &= \exp(aL\log 2) - 1 - L + \frac{1}{a\log 2}\exp(aL\log 2) - \frac{1}{a\log 2} \\ &\le \left(1+\frac{1}{a\log 2}\right)2^{aL}.\end{aligned}$$

We have used the substitutions $t = e^l$ and $u = t^{a\log 2}$. ■

Lemma 3: For any integer L, write $S_L = \{0, 1, \ldots, 2^{-L}\}$. Then the set $N_{L_q}^{(q)} = \{\mathbf{x} \in \mathbb{R}^{d_q} : \|\mathbf{x}\| \le 1, x_i^2 \in S_L\}$ forms a $(1/2)$-net for $\mathbb{S}^{d_q-1}$, $1 \le q \le p$ by taking $L_q = \log d_q + c_0$ for some constant c_0.

Proof: The result follows from Lemma 10 of [38] plugging in $\lambda = 1/d$. ■

Lemma 4: $|N_l^{(q)}| < |N_l^{(q)}| < \exp(C2^l(1+L-l))$.

Proof: The proof follows from Lemma 4 of [37]. In the notation of the proof of Lemma 4 from [37], we have that $N_{<l}^{(q)} = V_l^n$ where $l \le n$ and $n = \log(d_q)/\log(2)$. Then one has the required statement using the bound on $\log\#V_l^n$ from the proof of Lemma 4 and by noting that $n \le C\log(d_{\max}) = CL$. ■

References

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York, NY, USA: Wiley, 1984.

[2] I. Jolliffe, *Principal Component Analysis for Special Types of Data*. New York, NY, USA: Springer, 2002.

[3] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 14, pp. 1253–1278, 2000.

[4] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, Jan. 2008.

[5] T. Liu, M. Yuan, and H. Zhao, "Characterizing spatiotemporal transcriptome of the human brain via low-rank tensor decomposition," *Statist. Biosci.*, 2022, doi: 10.1007/s12561-021-09331-5.

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

[7] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.

[8] A. Cichocki *et al.*, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar. 2015.

[9] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, Jal. 2017.

[10] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, vol. 42. Berlin, Germany: Springer, 2012.

[11] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, p. 45, Nov. 2013.

[12] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[13] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2897–2905.

[14] S. B. Hopkins, J. Shi, and D. Steurer, "Tensor principal component analysis via sum-of-square proofs," in *Proc. 28th Conf. Learn. Theory*, 2015, pp. 956–1006.

[15] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, "Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors," in *Proc. 48th Annu. ACM Symp. Theory Comput.*, Jun. 2016, pp. 178–191.

[16] G. B. Arous, S. Mei, A. Montanari, and M. Nica, "The landscape of the spiked tensor model," *Commun. Pure Appl. Math.*, vol. 72, no. 11, pp. 2282–2330, Nov. 2019.

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

[18] A. Auddy and M. Yuan, "Perturbation bounds for (nearly) orthogonally decomposable tensors," 2020, *arXiv:2007.09024*.

[19] X. Wei and S. Minsker, "Estimation of the covariance structure of heavy-tailed distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[20] S. Minsker and X. Wei, "Robust modifications of U-statistics and applications to covariance estimation problems," *Bernoulli*, vol. 26, no. 1, pp. 694–727, Feb. 2020.

[21] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 28, pp. 7900–7905, Jul. 2016.

[22] E. Purdom and S. P. Holmes, "Error distribution for gene expression data," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, pp. 1–35, Jan. 2005.

[23] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, 2007, pp. 109–120.

[24] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin, "A note on the largest eigenvalue of a large dimensional sample covariance matrix," *J. Multivariate Anal.*, vol. 26, no. 2, pp. 166–168, Aug. 1988.

[25] Y. Seginer, "The expected norm of random matrices," *Combinat., Probab. Comput.*, vol. 9, no. 2, pp. 149–166, 2000.

[26] G. B. Arous and A. Guionnet, "The spectrum of heavy tailed random matrices," *Commun. Math. Phys.*, vol. 278, no. 3, pp. 715–751, Mar. 2008.

[27] A. Auffinger, G. Ben Arous, and S. Péché, "Poisson convergence for the largest eigenvalues of heavy tailed random matrices," *Annales de ĺIHP Probabilités et Statistiques*, vol. 45, no. 3, pp. 589–610, Aug. 2009.

[28] J. Ding, S. Hopkins, and D. Steurer, "Estimating rank-one spikes from heavy-tailed noise via self-avoiding walks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5576–5586.

[29] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Annales de ĺInstitut Henri Poincaré, Probabilités et Statistiques*, vol. 48, no. 4, pp. 1148–1185, Nov. 2012.

[30] O. Catoni, "PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design," 2016, *arXiv:1603.05229*.

[31] S. Mendelson and N. Zhivotovskiy, "Robust covariance estimation under $l_4 - l_2$ norm equivalence," *Ann. Statist.*, vol. 48, no. 3, pp. 1648–1664, 2020.

[32] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li, "Robust estimation of high-dimensional covariance and precision matrices," *Biometrika*, vol. 105, no. 2, pp. 271–284, Jun. 2018.

[33] I. Giulini, "PAC-Bayesian bounds for principal component analysis in Hilbert spaces," 2015, *arXiv:1511.06263*.

[34] S. Minsker, "Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries," *Ann. Statist.*, vol. 46, no. 6A, pp. 2871–2903, Dec. 2018.

[35] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, "User-friendly covariance estimation for heavy-tailed distributions," *Stat. Sci.*, vol. 34, no. 3, pp. 454–471, Aug. 2019.

[36] J. W. Silverstein, "On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix," *J. Multivariate Anal.*, vol. 30, no. 2, pp. 307–311, Aug. 1989.

[37] R. Latała, "Some estimates of norms of random matrices," *Proc. Amer. Math. Soc.*, vol. 133, no. 5, pp. 1273–1282, 2005.

[38] N. H. Nguyen, P. Drineas, and T. D. Tran, "Tensor sparsification via a bound on the spectral norm of random tensors," *Inf. Inference*, vol. 4, no. 3, pp. 195–229, 2015.

[39] T. Zhang and G. H. Golub, "Rank-one approximation to high order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 2, pp. 534–550, 2001.

[40] P. Raghavan, "Probabilistic construction of deterministic algorithms: Approximating packing integer programs," *J. Comput. Syst. Sci.*, vol. 37, no. 2, pp. 130–143, Oct. 1988.

[41] R. Latała, "Estimation of moments of sums of independent real random variables," *Ann. Probab.*, vol. 25, no. 3, pp. 1502–1513, Jul. 1997.

[42] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[43] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Berlin, Germany: Springer-Verlag, 2013.

[44] Z. D. Bai and Y. Q. Yin, "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix," *Ann. Probab.*, vol. 21, no. 3, pp. 108–127, Jul. 2008.

[45] R. Vershynin, *Introduction to the Non-Asymptotic Analysis of Random Matrices*. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 210–268.

Arnab Auddy received the master's degree in statistics from the Indian Statistical Institute, Kolkata, in 2018. He is currently pursuing the Ph.D. degree with the Statistics Department, Columbia University, where he is advised by Prof. Ming Yuan.

Ming Yuan received the Ph.D. degree in statistics from the University of Wisconsin–Madison in 2004. He has been a Professor of statistics at Columbia University since 2017. He was previously a Senior Investigator in virology at the Morgridge Institute for Research and a Professor of statistics at the University of Wisconsin–Madison, and prior to that a Coca-Cola Junior Professor of industrial and systems engineering at the Georgia Institute of Technology.