Effective Tensor Sketching via Sparsification

Dong Xia and Ming Yuan

***Abstract*— In this article, we investigate effective sketching schemes via sparsification for high dimensional multilinear arrays or tensors. More specifically, we propose a novel tensor sparsification algorithm that retains a subset of the entries of a tensor in a judicious way, and prove that it can attain a given level of approximation accuracy in terms of tensor spectral norm with a much smaller sample complexity when compared with existing approaches. In particular, we show that for a kth order $d \times \cdots \times d$ cubic tensor of *stable rank* r_s, the sample size requirement for achieving a relative error ε is, up to a logarithmic factor, of the order $r_s^{1/2} d^{k/2}/\varepsilon$ when ε is relatively large, and $r_s d/\varepsilon^2$ and essentially optimal when ε is sufficiently small. It is especially noteworthy that the sample size requirement for achieving a high accuracy is of an order independent of k. To further demonstrate the utility of our techniques, we also study how higher order singular value decomposition (HOSVD) of large tensors can be efficiently approximated via sparsification.**

***Index Terms*— Sketching, singular value decompostion (SVD), sparsification, tensor.**

I. Introduction

MASSIVE datasets are being generated everyday across diverse fields and can often be formatted into matrices or higher order tensors. For example, in biomedical research, huge data matrices and tensors arise in gene expression analysis [1], protein-to-protein interaction [2], and MRI image analysis [3]. They also occur frequently in statistical physics [4], [5], video processing [6], [7], and analyzing large graphs and social networks [8]–[10], to name a few. As the size of these data matrices or tensors grows, it becomes costly and sometimes prohibitively expensive to store, communicate or manipulate them. This naturally brings about the task of "sketching": approximate the original data matrices or tensors with a more manageable amount of sketches.

In the case of data matrices, numerous sketching approaches have been proposed in recent years. See [11] for a recent review. A popular idea behind many of these approaches is *sparsification* – creating a sparse matrix by zeroing out some entries of the original data matrix. Sparse sketching of a large data matrix not only reduces space complexity but also allows for efficient computations. See, e.g., [12]–[17], among others.

Manuscript received November 16, 2017; revised July 21, 2019; accepted August 28, 2020. Date of publication January 5, 2021; date of current version January 21, 2021. The work of Dong Xia was supported in part by Hong Kong RGC under Grant ECS 26302019. The work of Ming Yuan was supported in part by NSF under Grant DMS-1803450 and Grant DMS-2015285. *(Corresponding author: Ming Yuan.)*

Dong Xia is with the Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong (e-mail: madxia@ust.hk).

Ming Yuan is with the Department of Statistics, Columbia University, New York, NY 10027 USA (e-mail: ming.yuan@columbia.edu).

Communicated by P. Grohs, Associate Editor for Signal Processing.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3049174.

Digital Object Identifier 10.1109/TIT.2021.3049174

The main purpose of this article is to investigate to what extent sparsification can be used to effectively sketch higher order tensors. There have been some recent attempts along this direction. In particular, our work is inspired by [18] who showed that for a kth order cubic tensor $\mathbf{A} \in \mathbb{R}^{d\times\cdots\times d}$, there is a randomized sparsification scheme that yields another tensor $\tilde{\mathbf{A}}$ of same dimension but with

$$\mathrm{nnz}(\tilde{\mathbf{A}}) = \tilde{O}_p\left(\frac{d^{k/2}\mathrm{sr}(\mathbf{A})}{\varepsilon^2}\right), \qquad \text{as } d \to \infty, \tag{1}$$

such that

$$\|\tilde{\mathbf{A}} - \mathbf{A}\| \le \varepsilon\|\mathbf{A}\|.$$

Here, $\mathrm{nnz}(\cdot)$ stands for the number of nonzero entries of a tensor, $\mathrm{sr}(\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{F}}^2/\|\mathbf{A}\|^2$ is the so-called stable rank [16], [18] of a tensor $\mathbf{A}$, $\|\cdot\|$ is the usual tensor spectral norm, and $\tilde{O}(\cdot)$ means $O(\cdot)$, up to a certain polynomial of logarithmic factor. Note that the stable rank of a tensor can be viewed as a more stable alternative to the normal concept of ranks, as the name suggests. In particular, if a tensor is orthogonally decomposable, then its stable rank is always upper bounded by its rank and could be much smaller. In general, it can always be upper bounded by the multiplication of its smaller Tucker ranks. Similar results have also been obtained by [19] in the case when $k = 3$. On the one hand, the sample size requirement given by (1) is satisfying because it is essentially optimal in the matrix case, that is $k = 2$. See, e.g., [16]. On the other hand, the exponential dependence on k suggests a large amount of entries still need to be retained to yield a good approximation. Our goal is to investigate if this aspect could be improved.

In particular, we propose a novel tensor sparsification algorithm that randomly retain entries from $\mathbf{A}$ in a judicious way to yield a tensor $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ such that

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}} - \mathbf{A}\| \le \varepsilon\|\mathbf{A}\|,$$

and

$$\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}}) = \tilde{O}_p\left(\max\left\{\frac{d\cdot\mathrm{sr}(\mathbf{A})}{\varepsilon^2}, \frac{d^{k/2}\cdot\mathrm{sr}(\mathbf{A})^{1/2}}{\varepsilon}\right\}\right). \tag{2}$$

Here, to fix ideas, we focus on the case of cubic tensors although our results deal with more general rectangular tensors as well. This sample size requirement significantly improves those earlier ones. Especially if a high accuracy approximation is sought, that is $\varepsilon \le \mathrm{sr}(\mathbf{A})\cdot d^{-k/2+1}$, then our sparsification algorithm can achieve relative approximation error ε in terms of tensor spectral norm by retaining as few as $\tilde{O}_p(d\cdot\mathrm{sr}(\mathbf{A})\cdot\varepsilon^{-2})$ entries of $\mathbf{A}$. In addition, for approximation with lower precision, namely $\varepsilon > \mathrm{sr}(\mathbf{A})\cdot d^{-k/2+1}$, the number

of nonzero entries we keep is also substantially smaller than earlier proposals in that it depends on ε^{-1} *linearly* rather than quadratically, and on the stable rank $\mathrm{sr}(\mathbf{A})$ only through its square root.

Similar to other sparsification algorithms, we treat different entries according to their magnitude: large entries are always kept, and moderate ones are sampled proportion to their square values. The key difference between our approach and the existing ones is in the treatment of small entries. Instead of zeroing them out as, for example, [18], we sample them in a uniform fashion, which proves to be essential for obtaining good approximation with tighter number of nonzero entries. This modification is motivated by the concentration behavior of randomly sampled tensors recently observed by [20]–[22].

To demonstrate the effectiveness of our tensor sketching schemes, we show how they can be used for efficient approximation of the leading singular spaces from higher order singular value decomposition (HOSVD). Let $\mathbf{U}_j \in \mathbb{R}^{d\times r}$ be the top r left singular vectors of the flattening of $\mathbf{A}$ along its jth mode. We show that it is possible to construct an approximation $\widehat{\mathbf{U}}_j$ obeying

$$\|\widehat{\mathbf{U}}_j\widehat{\mathbf{U}}_j^\top - \mathbf{U}_j\mathbf{U}_j^\top\| \le \varepsilon,$$

if we retain

$$\tilde{O}_p\left(\max\left\{\frac{rd}{\varepsilon^2}, \frac{rd^{k/2}}{\varepsilon}\right\}\right)$$

carefully chosen entries. As before, we note that for high accuracy approximations, the sample complexity is essentially independent of the order of the tensor. Although our primary focus is on higher order tensors, as a byproduct, our results indicate that our sparsification scheme improves the sample complexity of earlier approaches for approximating the singular vectors of highly rectangular matrix.

The rest of the paper is organized as follows. We first discuss the new tensor sparsification algorithm in Section II. In Section III we consider the application to HOSVD. All proofs are relegated to Section IV.

II. Tensor Sparsification

Sketches of a tensor $\mathbf{A} \in \mathbb{R}^{d_1\times\ldots\times d_k}$ are its approximations, and we consider measuring their quality in terms of relative spectral norm. Recall that the spectral norm of a tensor $\mathbf{B} \in \mathbb{R}^{d_1\times\ldots\times d_k}$ is defined as

$$\|\mathbf{B}\| = \sup_{\mathbf{u}_j\in\mathbb{R}^{d_j}, \|\mathbf{u}_j\|_{\ell_2}\le 1} \langle \mathbf{B}, \mathbf{u}_1\otimes\ldots\otimes\mathbf{u}_k\rangle.$$

We seek an approximation $\widehat{\mathbf{A}}$ of $\mathbf{A}$ such that

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| \le \varepsilon\|\mathbf{A}\|,$$

for a prespecified accuracy $\varepsilon \in (0, 1)$. The error measure in terms of tensor spectral norm is common and especially ensures the approximation is suitable for downstream computation of low rank approximations.

The idea of sparsification is to systematically zero out entries of $\mathbf{A}$ and scale the remaining entries to yield a good approximation to it. We focus here on sparsification strategies that are carried out in an entry-by-entry fashion. Our approach can be characterized as *keeping* large entries, *sampling* moderate entries according to their magnitudes, and *sampling uniformly* small entries. The key is determining how to classify entries into these categories so that the number of nonzero entries retained are as small as possible. Details are presented in Algorithm 1.

In particular, we keep all entries whose absolute value is greater than $n^{-1/2}\|\mathbf{A}\|_{\mathrm{F}}$, sample uniformly all entries whose absolute value is smaller than $(d_1\cdots d_k)^{-1/2}\|\mathbf{A}\|_{\mathrm{F}}$, and sample proportional to their squared values entries whose absolute value is in-between. Here n is a sampling parameter. Note that $\mathbb{E}[\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}})] \le 2n$. And it is not hard to see, by Chernoff bound, that $\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}}) = O_p(n)$. In other words, n represents essentially the targeted sampling budget.

We note that our sparsification algorithm is similar to the one proposed earlier by [18]. But the two schemes also have several key differences. The main difference lies in the treatment of "small" entries. Reference [18] suggests to zero them out, while ours sample them in a uniform fashion. This is largely motivated by the concentration behavior of randomly sampled tensors observed earlier. In particular, it can be shown that a uniformly sampled tensor concentrates much sharply around its mean if its entries are sufficiently small [20]. Therefore, instead of discarding small entries, we could derive a good estimate of them by sampling uniformly. Another subtle difference between the two algorithm is in the criteria for "small" entries. Our criterion for "small" entries is that their absolute values are smaller than $(d_1\cdots d_k)^{-1/2}\|\mathbf{A}\|_{\mathrm{F}}$, whereas [18] treats only cubic tenors, that is $d_1 = d_2 = \cdots = d_k =: d$, and small entries of their scheme are those smaller than $n^{-1/2}d^{-k/4}\|\mathbf{A}\|_{\mathrm{F}}\log^{k/2} d$.

We now present the performance bounds for our sparsification algorithm.

Theorem 1: Let $\mathbf{A} \in \mathbb{R}^{d_1\times\ldots\times d_k}$ and $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ be the output from Algorithm 1 with sampling budget n. There exists an absolute constant $C > 0$ such that if for any $\alpha \ge 4\log(k\log d_{\max})$ and $\varepsilon \in (0, 1)$, if

$$n \ge C\max\left\{\alpha^4 k^8 \frac{d_{\max}\cdot\mathrm{sr}(\mathbf{A})}{\varepsilon^2}\log^2 d_{\max}, \alpha^2\ k^5\frac{(d_1\cdots d_k\cdot\mathrm{sr}(\mathbf{A}))^{1/2}}{\varepsilon}\log^{k+4} d_{\max}\right\},$$

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

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}} - \mathbf{A}\| \le \varepsilon\|\mathbf{A}\|,$$

where $d_{\max} = \max\{d_1, \ldots, d_k\}$.

In the light of Theorem 1, we can achieve relative error ε in terms of tensor spectral norm with a sparse tensor such that

$$\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}}) = \begin{cases}\tilde{O}\left(\varepsilon^{-2}d_{\max}\cdot\mathrm{sr}(\mathbf{A})\right), & \text{if } \varepsilon \le \frac{d_{\max}\cdot\mathrm{sr}(\mathbf{A})^{1/2}}{(d_1\ldots d_k)^{1/2}}\\ \tilde{O}\left(\varepsilon^{-1}(d_1\ldots d_k\cdot\mathrm{sr}(\mathbf{A}))^{1/2}\right), & \text{otherwise}\end{cases}$$

This significant improves earlier work by [19] and [18]. It is worth noting that for small ε, or high accuracy approximation, the number of nonzero entries of $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ is of the order $\varepsilon^{-2}d_{\max}\cdot\mathrm{sr}(\mathbf{A})$. This, in particular, is known to be optimal in the matrix ($k = 2$) case [16]. On the other hand, for lower accuracy

Algorithm 1 Tensor Sparsification

Input: $\mathbf{A} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$, sampling budget $n \in [1, d_1 \cdots d_k]$.
Output: $\widehat{\mathbf{A}}^{\mathrm{SPA}} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$.
for $i_1 \in [d_1], i_2 \in [d_2], \ldots, i_k \in [d_k]$ **do**
  if $|A(i_1, \ldots, i_k)| \geq \|\mathbf{A}\|_{\mathrm{F}}/n^{1/2}$, **then**
    $\widehat{A}(i_1, \ldots, i_k) = A(i_1, \ldots, i_k)$.
  end if
  if $|A(i_1, \ldots, i_k)|/\|\mathbf{A}\|_{\mathrm{F}} \in \left(\frac{1}{(d_1 \cdots d_k)^{1/2}}, \frac{1}{n^{1/2}}\right)$, **then**

$$\widehat{A}(i_1, \ldots, i_k) = \begin{cases} \frac{A(i_1, \ldots, i_k)}{P(i_1, \ldots, i_k)}, & \text{with probability } P(i_1, \ldots, i_k) := \frac{nA^2(i_1, \ldots, i_k)}{\|\mathbf{A}\|_{\mathrm{F}}^2} \\ 0, & \text{with probability } 1 - P(i_1, \ldots, i_k). \end{cases}$$

  end if
  if $|A(i_1, \ldots, i_k)| \leq \|\mathbf{A}\|_{\mathrm{F}}/(d_1 \ldots d_k)^{1/2}$, **then**

$$\widehat{A}(i_1, \ldots, i_k) = \begin{cases} \frac{A(i_1, \ldots, i_k)}{P(i_1, \ldots, i_k)}, & \text{with probability } P(i_1, \ldots, i_k) := \frac{n}{d_1 d_2 \cdots d_k} \\ 0, & \text{with probability } 1 - P(i_1, \ldots, i_k) \end{cases}$$

  end if
end for
Output: $\widehat{\mathbf{A}}^{\mathrm{SPA}} = \widehat{\mathbf{A}}$.

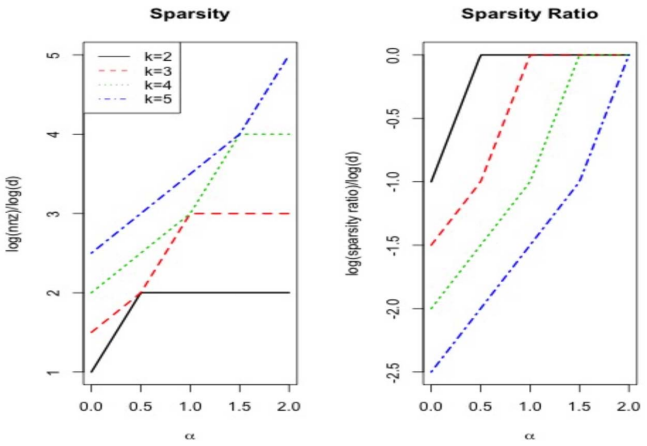


Fig. 1. Sparsity or sparsity ratio of the proposed sparse approximation to a cubic tensor versus its accuracy.

approximation, the dependence of $\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}})$ on ε^{-1} is linear rather than quadratic as existing sparsification schemes.

To better appreciate the bounds given in Theorem 1, it is instructive to consider cubic tensors, that is $d_1 = \cdots = d_k = d$, and the case when $\mathrm{sr}(\mathbf{A}) = O(1)$ and $\varepsilon = d^{-\alpha}$ for some $\alpha > 0$. Theorem 1 indicates that $\mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}}) = \tilde{O}(d^{\min\{k, \max\{1+2\alpha, k/2+\alpha\}\}})$. Note that a kth order cubic tensor has d^k entries so that the sparsity ratio of $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ is $d^{-k} \cdot \mathrm{nnz}(\widehat{\mathbf{A}}^{\mathrm{SPA}}) = \tilde{O}(d^{\min\{0, \max\{1+2\alpha-k, \alpha-k/2\}\}})$. In particular, Figure 1 plots the exponent of such sparisity bounds versus the exponent of accuracy α for $k = 2, 3, 4$ and 5.

The main technical tool for proving Theorem 1 is the following concentration inequality for random tensors which might be of independent interest.

Theorem 2: Let $\mathbf{A} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$ and $\mathbf{P} \in [0, 1]^{d_1 \times \ldots \times d_k}$ be two fixed tensors, $\mathbf{\Delta} \in \{0, 1\}^{d_1 \times \ldots \times d_k}$ be a random tensor such that $\mathbb{E}\Delta(i_1, \ldots, i_k) = P(i_1, \ldots, i_k)$. Define a random tensor $\widehat{\mathbf{A}} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$ by

$$\widehat{A}(i_1, \ldots, i_k) = A(i_1, \ldots, i_k)\Delta(i_1, \ldots, i_k)/P(i_1, \ldots, i_k).$$

Then, there exist absolute constants $C_1, C_2, C_3 > 0$ such that for any $\alpha > 0$, with probability at least $1 - 3d_{\max}^{-\alpha}$,

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| \leq C_1\Big(\Big(\sum_{j=1}^{k} d_j\Big)^{1/2} + \alpha k \log d_{\max}\Big)\alpha_{2,\infty}(\mathbf{A}, \mathbf{P}) + C_2\alpha k^3 \log^{k+2}(d_{\max})\sqrt{\nu}\alpha_{\infty}(\mathbf{A}, \mathbf{P}),$$

where

$$\nu = C_3\alpha \max\big\{\beta(\mathbf{P}), k \log d_{\max}\big\},$$

$$\beta(\mathbf{P}) = \max_{j=1,\ldots,k} \max_{i_1,\ldots,i_{j-1},i_{j+1},\ldots,i_k} \sum_{i_j=1}^{d_j} P(i_1, \ldots, i_k),$$

$$\alpha_{\infty}(\mathbf{A}, \mathbf{P}) = \max_{i_j \in [d_j], j=1,\ldots,k} \frac{|A(i_1, \ldots, i_k)|}{P(i_1, \ldots, i_k)},$$

and

$$\alpha_{2,\infty}(\mathbf{A}, \mathbf{P}) = \max_{i_j \in [d_j], j=1,\ldots,k} \left(\frac{A^2(i_1, \ldots, i_k)}{P(i_1, \ldots, i_k)}\right)^{1/2}.$$

Here we follow the convention that $0/0 = 0$.

III. Effective Computation of HOSVD via Sparsification

To further illustrate the merits of the sketching schemes introduced earlier, we now consider a specific application to HOSVD, a popular technique for analyzing high dimensional tensor data. See, e.g., [23], [24] and references therein.

For a k-th order tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$, let $\mathbf{M}_j = \mathcal{M}_j(\mathbf{A}) \in \mathbb{R}^{d_j \times d_{-j}}$ be its j-th matricization where $1 \le j \le k$, that is,

$$\mathcal{M}_j(\mathbf{A})\Big(i_j, \sum_{s=1, s\neq j}^{k} (i_s - 1)\Big(\prod_{s'=s+1, s'\neq j}^{k} d_{s'}\Big) + 1\Big) = A(i_1, \ldots, i_k), \quad \forall i_j \in [d_j], 1 \le j \le k.$$

Here $d_{-j} = (d_1 \cdots d_k)/d_j$. Denote by $\mathbf{U}_j^{(r_j)}$ the collection of the top r_j left singular vectors of $\mathbf{M}_j$. Clearly, $\mathbf{U}_j^{(r_j)}$ is computable via the standard matrix singular value decomposition on $\mathbf{M}_j$ whose computation complexity is $O(d_j d_1 d_2 \ldots d_k)$, see [25]. Efficient computation of singular value decomposition for large matrices is an actively researched topic in numerical algebra and computational science. See [14], [15], [26]–[29], among numerous others.

A general idea is to first obtain an approximation of $\mathbf{M}_j$, say $\widehat{\mathbf{M}}_j \in \mathbb{R}^{d_j \times d_{-j}}$, that is amenable for fast computation of singular value decomposition; and then approximate $\mathbf{U}_j^{(r_j)}$ by the top left singular vectors of $\widehat{\mathbf{M}}_j$. In particular, sparsification is a powerful tool for fast computation of singular vectors. See, e.g., [11], [14].

Denote by $\mathbf{\Delta}_j = \widehat{\mathbf{M}}_j - \mathbf{M}_j$ and by $\widehat{\mathbf{U}}_j^{(r_j)}$ the leading r_j left singular vectors of $\widehat{\mathbf{M}}_j$. By Davis-Kahan Theorem [30], we get

$$\big\|\widehat{\mathbf{U}}_j^{(r_j)}\big(\widehat{\mathbf{U}}_j^{(r_j)}\big)^\top - \mathbf{U}_j^{(r_j)}\big(\mathbf{U}_j^{(r_j)}\big)^\top\big\| \le \frac{2\|\mathbf{\Delta}_j\|}{\bar{g}_{r_j}(\mathbf{M}_j)} \tag{3}$$

where $\sigma_k(\cdot)$ denotes the k-th singular value, and

$$\bar{g}_{r_j}(\mathbf{M}_j) = \sigma_{r_j}(\mathbf{M}_j) - \sigma_{r_j+1}(\mathbf{M}_j),$$

is the r_j-th eigengap. In particular, we can consider applying this strategy by taking $\widehat{\mathbf{M}}_j = \mathcal{M}_j(\widehat{\mathbf{A}}^{\mathrm{SPA}})$. The following result characterizes its performance.

Theorem 3: Let $\mathbf{U}_j^{(r_j)}$ and $\widehat{\mathbf{U}}_j^{(r_j)}$ be the top r_j left singular vectors of $\mathcal{M}_j(\mathbf{A})$ and $\mathcal{M}_j(\widehat{\mathbf{A}}^{\mathrm{SPA}})$ respectively. Then there exists an absolute constant $C > 0$ such that for any $t > 0$,

$$\big\|\widehat{\mathbf{U}}_j^{(r_j)}\big(\widehat{\mathbf{U}}_j^{(r_j)}\big)^\top - \mathbf{U}_j^{(r_j)}\big(\mathbf{U}_j^{(r_j)}\big)^\top\big\| \le C\frac{\|\mathbf{M}_j\|_{\mathrm{F}}}{\bar{g}_{r_j}(\mathbf{M}_j)}\Bigg(\sqrt{\frac{d_1 \ldots d_k(t + k\log d_{\max})}{n d_j}} + \frac{(d_1 \ldots d_k)^{1/2}(t + k\log d_{\max})}{n}\Bigg),$$

with probability at least $1 - e^{-t}$.

By Theorem 3, in the case when $\|\mathbf{M}_j\|_{\mathrm{F}}/\bar{g}_{r_j}(\mathbf{M}_j) = O(\sqrt{r_j})$, we can ensure

$$\big\|\widehat{\mathbf{U}}_j^{(r_j)}\big(\widehat{\mathbf{U}}_j^{(r_j)}\big)^\top - \mathbf{U}_j^{(r_j)}\big(\mathbf{U}_j^{(r_j)}\big)^\top\big\| \le \varepsilon$$

by taking

$$n \ge C \cdot \max\left\{\frac{r_j d_1 \ldots d_k}{d_j \varepsilon^2}, \frac{(r_j d_1 \ldots d_k)^{1/2}}{\varepsilon}\right\}\log d_{\max}. \tag{4}$$

A critical fact that is neglected by this approach is that we are interested in approximating the left singular vectors of a potentially very "fat" matrix because d_{-j} is generally much larger than d_j. As such, this type of approach turns out to be suboptimal for our purpose.

Alternatively, we adopt a new spectral method similar in spirit to a recent proposal from [22]. More specifically, we shall approximate $\mathbf{U}_j^{(r_j)}$ by the leading eigenvectors of an approximation of $\mathbf{M}_j\mathbf{M}_j^\top$ instead. In particular, we can run Algorithm 1 twice to obtain two *independent* sparsifications of $\mathbf{A}$, denoted by $\widehat{\mathbf{A}}_1^{\mathrm{SPA}}$ and $\widehat{\mathbf{A}}_2^{\mathrm{SPA}}$, and then proceed to approximate $\mathbf{M}_j\mathbf{M}_j^\top$ by $\mathcal{M}_j(\widehat{\mathbf{A}}_1^{\mathrm{SPA}})\mathcal{M}_j(\widehat{\mathbf{A}}_2^{\mathrm{SPA}})^\top$. Details are presented in Algorithm 2.

Algorithm 2 Computing HOSVD via Tensor Sparsification

Input: $\mathbf{A} \in \mathbb{R}^{d_1 \times \ldots \times d_k}$, sampling budget $n \ge 1$.
Output: the r_j leading left singular vectors $\widehat{\mathbf{U}}_j^{(r_j)}$ as an estimate of HOSVD of $\mathcal{M}_j(\mathbf{A})$.
Run Algorithm 1 on $\mathbf{A}$ with sampling budget n. Denote the output by $\widehat{\mathbf{A}}_1^{\mathrm{SPA}}$.
Run Algorithm 1 on $\mathbf{A}$ with sampling budget n. Denote the output by $\widehat{\mathbf{A}}_2^{\mathrm{SPA}}$.
Compute $\widehat{\mathbf{U}}_j^{(r_j)}$ as the r_j leading left singular vectors of $\mathcal{M}_j(\widehat{\mathbf{A}}_1^{\mathrm{SPA}})\mathcal{M}_j(\widehat{\mathbf{A}}_2^{\mathrm{SPA}})^\top$.
Output $\widehat{\mathbf{U}}_j^{(r_j)}$.

The following theorem provides the performance bound for approximate the singular space $\mathbf{U}_j^{(r_j)}$s.

Theorem 4: Denote by $\mathbf{U}_j^{(r_j)}$ the r_j leading left singular vectors of $\mathcal{M}_j(\mathbf{A})$. Let $\widehat{\mathbf{U}}_j^{(r_j)}$ be the output from Algorithm 2. There exists an absolute constant $C > 0$ such that for any $\alpha \ge 1$ and $\varepsilon \in (0, 1)$, if

$$n \ge C\alpha\Bigg(\frac{k^2 d_j \log d_{\max}}{\varepsilon^2}\frac{\|\mathbf{A}\|_{\mathrm{F}}^2\sigma_{\max}^2(\mathbf{M}_j)}{\bar{g}_{r_j}^2(\mathbf{M}_j\mathbf{M}_j^\top)} + \frac{k(d_1 \ldots d_k)^{1/2}\log d_{\max}}{\varepsilon}\frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{\bar{g}_{r_j}(\mathbf{M}_j\mathbf{M}_j^\top)}\Bigg),$$

then

$$\big\|\widehat{\mathbf{U}}_j^{(r_j)}\big(\widehat{\mathbf{U}}_j^{(r_j)}\big)^\top - \mathbf{U}_j^{(r_j)}\big(\mathbf{U}_j^{(r_j)}\big)^\top\big\| \le \varepsilon,$$

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

From Theorem 4, if $\|\mathbf{A}\|_{\mathrm{F}}^2/\bar{g}_{r_j}(\mathbf{M}_j\mathbf{M}_j^\top) = O(r_j)$ and $\|\mathbf{A}\|_{\mathrm{F}}^2\sigma_{\max}^2(\mathbf{M}_j)/\bar{g}_{r_j}^2(\mathbf{M}_j\mathbf{M}_j^\top) = O(r_j)$, then the required sample complexity for sparsification is

$$\tilde{O}_p\Bigg(\frac{k^2 r_j d_j \log d_{\max}}{\varepsilon^2} + \frac{k r_j (d_1 \ldots d_k)^{1/2}\log d_{\max}}{\varepsilon}\Bigg).$$

It is worth noting that, even though our main focus is on higher order tensors, in the case of matrices ($k = 2$) this sample complexity compares favorable with other sparsification techniques that have been developed for computing singular vectors. For example, consider computing the top r left singular vectors of a $d_1 \times d_2$ ($d_1 \le d_2$) matrix. The approach from [14] needs to sample

$$\tilde{O}_p\Big(\frac{r d_1 d_2^2}{\varepsilon^2} \cdot \frac{\max_{i,j}|A(i,j)|^2}{\|\mathbf{A}\|_{\mathrm{F}}^2}\Big)$$

entries; the technique of [31] requires

$$\tilde{O}_p\Big(\frac{rd_2}{\varepsilon^2}\Big)$$

entries. These are to be compared with Algorithm 2 which needs

$$\tilde{O}_p\Big(\frac{rd_1}{\varepsilon^2}+\frac{r(d_1d_2)^{1/2}}{\varepsilon}\Big)$$

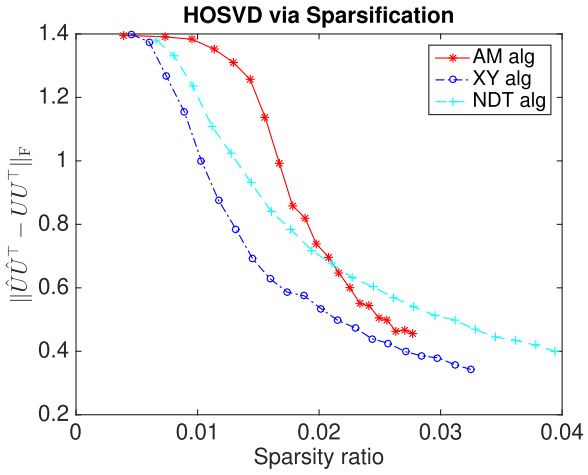
sampled entries, which could be much smaller than the previous two when $d_1 \ll d_2$.

To further illustrate the practical merits of our sparsification schemes in computing HOSVD, we conducted a small numerical experiment comparing our algorithm with a couple of notable alternatives developed earlier by [14] and [18] respectively. More specifically, we generated a tensor $\mathbf{A} \in \mathbb{R}^{d\times d\times d}$ as follows:

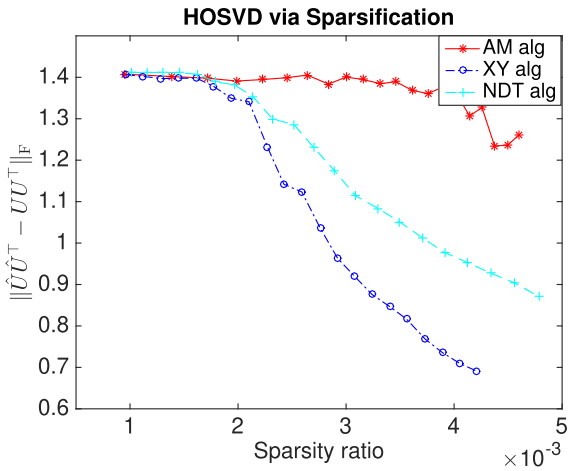
$$\mathbf{A}=\sum_{i=1}^{5}(\mathbf{a}_i\otimes\mathbf{b}_i\otimes\mathbf{c}_i)+\mathbf{Z} \tag{5}$$

where $\mathbf{a}_i,\mathbf{b}_i,\mathbf{c}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}_d)$, and $\mathbf{Z}$ has independent centered Gaussian entries. To introduce heterogeneity among the entries, we set $Z(i_1,i_2,i_3)\sim\mathcal{N}\big(0,1/\log((i_1-1)d^2+(i_2-1)d+i_3+1)\big)$ for all $i_1,i_2,i_3\in[d]$. We considered in particular $d=100$ or 200. In each simulatin run, we applied the three sparsification algorithms to approximately compute the top-r left singular vectors of $\mathcal{M}_1(\mathbf{A})$. We report the averaged loss, $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top-\mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}$, based on 20 simulation runs versus the sparsity ratio, $\mathrm{nnz}(\widehat{\mathbf{A}})/d^3$, for the three algorithms in Figure 2. Here $\mathbf{U}$ and $\widehat{\mathbf{U}}$ denote the top-r left singular vectors of the $\mathcal{M}_1(\mathbf{A})$ and its sparse approximations respectively.

It is clear from Figure 2 that our algorithms yields much sparser approximations at the same level of accuracy as the algorithms from [14] and [18] in this instance. It is also noteworthy that such advantage becomes clearer as the size of tensors increases which makes our algorithm more suitable for large scale applications.



(a) $d=100$



(b) $d=200$

Fig. 2. "XY alg", "AM alg" and "NDT alg" correspond to Algorithm 2, and the algorithms from [14] and [18] respectively.

IV. Proofs

We now present the proofs to our main results.

A. Proof of Theorem 1

Theorem 1 follows immediately from the concentration bound for $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}-\mathbf{A}\|$ below.

Lemma 1: Let $\mathbf{A}\in\mathbb{R}^{d_1\times\ldots\times d_k}$ and $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ be the output from Algorithm 1 with sampling budget n. Then there exist absolute constants $C_1,C_2>0$ such that, for any $\alpha\geq 4\log(k\log d_{\max})$, the following bound holds with probability at least $1-d_{\max}^{-\alpha}$:

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}-\mathbf{A}\|\leq C_1\alpha^2k^4\log(d_{\max})\sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2d_{\max}}{n}} + C_2\alpha^2k^5\log^{k+4}(d_{\max})\frac{(d_1\ldots d_k)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}}{n}.$$

Proof of Lemma 1: Given $\mathbf{A}$, we define the disjoint subsets of $[d_1]\times\ldots\times[d_k]$

$$\Omega_1=\big\{(i_1,\ldots,i_k):|A(i_1,\ldots,i_k)|\leq\|\mathbf{A}\|_{\mathrm{F}}/(d_1\ldots d_k)^{1/2}\big\},$$

$$\Omega_2=\Big\{(i_1,\ldots,i_k):\frac{|A(i_1,\ldots,i_k)|}{\|\mathbf{A}\|_{\mathrm{F}}}\in\Big(\frac{1}{\sqrt{d_1\ldots d_k}},\frac{1}{n^{1/2}}\Big)\Big\},$$

and

$$\Omega_3=\big\{(i_1,\ldots,i_k):|A(i_1,\ldots,i_k)|\geq\|\mathbf{A}\|_{\mathrm{F}}/n^{1/2}\big\}.$$

Note that $\Omega_1,\Omega_2,\Omega_3$ are non-random subsets for given $\mathbf{A}$. Then,

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}-\mathbf{A}\|\leq\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_1}-\mathbf{A}_{\Omega_1}\|+\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2}-\mathbf{A}_{\Omega_2}\| +\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_3}-\mathbf{A}_{\Omega_3}\|.$$

By definition of $\widehat{\mathbf{A}}^{\mathrm{SPA}}$ in Algorithm 1, we have $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_3}-\mathbf{A}_{\Omega_3}\|=0$ so that it suffices to bound $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_1}-\mathbf{A}_{\Omega_1}\|$ and $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2}-\mathbf{A}_{\Omega_2}\|$.

Step 1: upper bound of $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_1}-\mathbf{A}_{\Omega_1}\|$. In order to apply Theorem 2, we introduce auxiliary tensors $\mathbf{B}$ and $\widetilde{\mathbf{B}}$ such that $\mathbf{B}_{\Omega_1}=\mathbf{A}_{\Omega_1}$ and $\mathbf{B}_{\Omega_1^\dagger}=\mathbf{0}$, where $\Omega_1^\dagger$ denotes the complement of Ω_1. Define a tensor $\mathbf{P}\in[0,1]^{d_1\times\ldots\times d_k}$ such that

$$P(i_1,\ldots,i_k)=\begin{cases}\frac{n}{d_1\ldots d_k}, & \text{if } (i_1,\ldots,i_k)\in\Omega_1\\ 0, & \text{otherwise.}\end{cases}$$

Then, random tensor $\widetilde{\mathbf{B}}$ is defined as

$$\widetilde{B}(i_1,\ldots,i_k)=\begin{cases}\frac{B(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}, & \text{with prob. } P(i_1,\ldots,i_k)\\ 0, & \text{with prob. } 1-P(i_1,\ldots,i_k),\end{cases}$$

where we followed the convention $0/0 = 0$. Clearly, $\widetilde{\mathbf{B}} - \mathbf{B}$ has the same distribution as $\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_1} - \mathbf{A}_{\Omega_1}$. To apply Theorem 2, we observe that

$$\nu = C_3 t \max\left\{\frac{n d_{\max}}{d_1 \dots d_k}, k \log d_{\max}\right\}$$

and

$$\alpha_\infty(\mathbf{B},\mathbf{P}) = \max_{i_1,\dots,i_k} \frac{|B(i_1,\dots,i_k)|}{P(i_1,\dots,i_k)} = \max_{i_1,\dots,i_k} \frac{d_1\dots d_k}{n}|B(i_1,\dots,i_k)| \le \frac{(d_1\dots d_k)^{1/2}}{n}\|\mathbf{A}\|_{\mathrm{F}}$$

and

$$\alpha_{2,\infty}(\mathbf{B},\mathbf{P}) = \max_{i_1,\dots,i_k} \frac{|B(i_1,\dots,i_k)|}{\sqrt{P(i_1,\dots,i_k)}} = \max_{i_1,\dots,i_k} \frac{(d_1\dots d_k)^{1/2}|B(i_1,\dots,i_k)|}{n^{1/2}} \le \frac{\|\mathbf{A}\|_{\mathrm{F}}}{n^{1/2}}.$$

By Theorem 2, with probability at least $1 - d_{\max}^{-t}$,

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_1} - \mathbf{A}_{\Omega_1}\| = \|\widetilde{\mathbf{B}} - \mathbf{B}\| \le C_1 t k^3 \sqrt{\frac{d_{\max}}{n}}\|\mathbf{A}\|_{\mathrm{F}} + C_2 t k^4 \log^{k+3}(d_{\max})\frac{(d_1\dots d_k)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}}{n}.$$

Step 2: upper bound of $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2} - \mathbf{A}_{\Omega_2}\|$. Bounding $\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2} - \mathbf{A}_{\Omega_2}\|$ is more involved. For $s = 1, 2, \dots, \lceil \log(d_1 \dots d_k/n)\rceil$, define

$$\Omega_{2,s} = \left\{(i_1,\dots,i_k) : |A(i_1,\dots,i_k)|^2 \in \Big[\frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}2^{-s}, \frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}2^{-s+1}\Big)\right\}.$$

Clearly,

$$\Omega_2 = \bigcup_{s=1}^{\lceil \log(d_1\dots d_k/n)\rceil} \Omega_{2,s},$$

so that

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2} - \mathbf{A}_{\Omega_2}\| \le \sum_{s=1}^{\lceil \log(d_1\dots d_k/n)\rceil} \big\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_{2,s}} - \mathbf{A}_{\Omega_{2,s}}\big\|.$$

We now apply Theorem 2 to bound each term on the righthand side. We follow the same strategy as before and define auxiliary tensors $\widetilde{\mathbf{B}}_s$ and $\mathbf{B}_s$ such that $\big(\mathbf{B}_s\big)_{\Omega_{2,s}} = \mathbf{A}_{\Omega_{2,s}}$ and $\big(\mathbf{B}_s\big)_{\Omega_{2,s}^\dagger} = \mathbf{0}$. The probability tensor $\mathbf{P}_s$ is defined as

$$P_s(i_1,\dots,i_k) = \begin{cases} \frac{nA^2(i_1,\dots,i_k)}{\|\mathbf{A}\|_{\mathrm{F}}^2}, & \text{if } (i_1,\dots,i_k) \in \Omega_{2,s} \\ 0, & \text{otherwise.}\end{cases}$$

The random tensor $\widetilde{\mathbf{B}}_s$ is defined as

$$\widetilde{B}_s(i_1,\dots,i_k) = \begin{cases} \frac{B_s(i_1,\dots,i_k)}{P_s(i_1,\dots,i_k)}, & \text{with prob. } P_s(i_1,\dots,i_k) \\ 0, & \text{with prob. } 1 - P_s(i_1,\dots,i_k).\end{cases}$$

Clearly, $\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_{2,s}} - \mathbf{A}_{\Omega_{2,s}}$ has the same distribution as $\widetilde{\mathbf{B}}_s - \mathbf{B}_s$. To apply Theorem 2, observe that

$$\alpha_{2,\infty}(\mathbf{B}_s,\mathbf{P}_s) = \max_{(i_1,\dots,i_k)\in\Omega_{2,s}} \sqrt{\frac{B_s^2(i_1,\dots,i_k)}{P_s(i_1,\dots,i_k)}} = \max_{(i_1,\dots,i_k)\in\Omega_{2,s}} \sqrt{\frac{A^2(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}} = \sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}}.$$

Since

$$\nu = C_1 t \max\left\{\max_{j\in[k]}\max_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_k}\sum_{i_j:(i_1,\dots,i_j)\in\Omega_{2,s}} P(i_1,\dots,i_k), k\log d_{\max}\right\},$$

we obtain

$$\begin{aligned}\sqrt{\nu}\alpha_\infty(\mathbf{B}_s,\mathbf{P}_s) &\le C_1 t^{1/2}k^{1/2}\log^{1/2}(d_{\max})\max_{(i_1,\dots,i_k)\in\Omega_{2,s}}\frac{|A(i_1,\dots,i_k)|}{P(i_1,\dots,i_k)} \\ &+ C_2 t^{1/2}\Bigg(\max_{j\in[k]}\max_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_k}\sqrt{\sum_{i_j:(i_1,\dots,i_k)\in\Omega_{2,s}} P(i_1,\dots,i_k)}\Bigg) \\ &\quad\cdot \max_{(i_1,\dots,i_k)\in\Omega_{2,s}}\frac{|A(i_1,\dots,i_k)|}{P(i_1,\dots,i_k)}.\end{aligned}$$

By definition of $\Omega_{2,s}$, we have

$$\frac{\max_{(i_1,\dots,i_k)\in\Omega_{2,s}} P(i_1,\dots,i_k)}{\min_{(i_1,\dots,i_k)\in\Omega_{2,s}} P(i_1,\dots,i_k)} \le 2.$$

Therefore,

$$\begin{aligned}&\Bigg(\max_{j\in[k]}\max_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_k}\sqrt{\sum_{i_j:(i_1,\dots,i_k)\in\Omega_{2,s}} P(i_1,\dots,i_k)}\Bigg) \\ &\qquad\cdot \max_{(i_1,\dots,i_k)\in\Omega_{2,s}}\frac{|A(i_1,\dots,i_k)|}{P(i_1,\dots,i_k)} \\ &\le \sqrt{2d_{\max}}\max_{(i_1,\dots,i_k)\in\Omega_{2,s}}\frac{|A(i_1,\dots,i_k)|}{\sqrt{P(i_1,\dots,i_k)}}.\end{aligned}$$

By the fact $P(i_1,\dots,i_k) = \frac{nA^2(i_1,\dots,i_k)}{\|\mathbf{A}\|_{\mathrm{F}}^2}$ and $|A(i_1,\dots,i_k)| \ge \|\mathbf{A}\|_{\mathrm{F}}/(d_1\dots d_k)^{1/2}$, we get

$$\begin{aligned}\sqrt{\nu}\alpha_\infty(\mathbf{B}_s,\mathbf{P}_s) &\le C_1 k^{1/2}t^{1/2}\log^{1/2}(d_{\max})\max_{(i_1,\dots,i_k)\in\Omega_{2,s}}\frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n|A(i_1,\dots,i_k)|} \\ &+ C_2 t^{1/2}d_{\max}^{1/2}\sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}} \\ &\le C_1 k^{1/2}t^{1/2}\log^{1/2}(d_{\max})\frac{(d_1\dots d_k)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}}{n} \\ &+ C_2 t^{1/2}\sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2 d_{\max}}{n}}.\end{aligned}$$

By Theorem 2, with probability at least $1 - d_{\max}^{-t}$,

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_{2,s}} - \mathbf{A}_{\Omega_{2,s}}\| \le C_1 t^2 k^3\sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2 d_{\max}}{n}} + C_2\, t^2 k^4 \log^{k+3}(d_{\max})\frac{(d_1\dots d_k)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}}{n}.$$

By taking a uniform bound for all $s = 1, 2, \dots, \lceil \log(d_1\dots d_k/n)\rceil$, we conclude that with probability at least $1 - k\log(d_{\max})d_{\max}^{-t}$,

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}}_{\Omega_2} - \mathbf{A}_{\Omega_2}\| \le C_1 t^2 k^4 \log(d_{\max})\sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2 d_{\max}}{n}} + C_2\, t^2 k^5 \log^{k+4}(d_{\max})\frac{(d_1\dots d_k)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}}{n}.$$

Final step: finalize the proof of Lemma 1. Put the above bounds together, we end up with, for any $t > 1$,

$$\|\widehat{\mathbf{A}}^{\mathrm{SPA}} - \mathbf{A}\| \le C_1 t^2 k^4 \log(d_{\max}) \sqrt{\frac{\|\mathbf{A}\|_{\mathrm{F}}^2 d_{\max}}{n}} + C_2\, t^2 k^5 \log^{k+4}(d_{\max}) \frac{(d_1 \dots d_k)^{1/2} \|\mathbf{A}\|_{\mathrm{F}}}{n}$$

which holds with probability at least $1 - \big(1 + k \log d_{\max}\big) d_{\max}^{-t} = 1 - d_{\max}^{-t + \log(k \log d_{\max})}$. □

B. Proof of Theorem 2

We begin with symmetrization [20] and obtain for any $t > 0$,

$$\mathbb{P}\Big(\|\widehat{\mathbf{A}} - \mathbf{A}\| \ge t\Big) \le 4\mathbb{P}\Big(\|\boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\| \ge 2t\Big) + 4 \exp\Big(\frac{-t^2/2}{\alpha_{2,\infty}^2(\mathbf{A}, \mathbf{P}) + t\alpha_\infty(\mathbf{A}, \mathbf{P})/3}\Big)$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^{d_1 \times \dots \times d_k}$ is a random tensor with i.i.d. Rademacher entries, and

$$\alpha_\infty(\mathbf{A}, \mathbf{P}) = \max_{i_j \in [d_j], j = 1, \dots, k} \frac{A(i_1, \dots, i_k)}{P(i_1, \dots, i_k)}$$

and

$$\alpha_{2,\infty}(\mathbf{A}, \mathbf{P}) = \max_{i_j \in [d_j], j \in [k]} \left(\frac{A^2(i_1, \dots, i_k)}{P(i_1, \dots, i_k)}\right)^{1/2}.$$

The $\odot$ operator stands for entrywse multiplication, that is

$$\big(\varepsilon \odot \widehat{A}\big)(i_1, \dots, i_k) = \varepsilon(i_1, \dots, i_k) \widehat{A}(i_1, \dots, i_k).$$

By definition, the operator norm $\|\boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\|$ is given by

$$\|\boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\| = \sup_{\mathbf{u}_j \in \mathbb{R}^{d_j}, \|\mathbf{u}_j\|_{\ell_2} \le 1, 1 \le j \le k} \big\langle \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}, \mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k \big\rangle.$$

We begin with the discretization of ℓ_2-norm balls. For each $j = 1, \dots, k$, define

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

where $m_j = 2\big(\lceil \log_2 d_j \rceil + 3\big)$. Define the "digitalization" operator $\mathbf{D}_s$ which zeros out the entries of $\mathbf{A}$ whose absolute value is not $2^{-s/2}$. Then,

$$\mathbf{D}_s(\mathbf{A}) = \sum_{i_1, \dots, i_k} \mathbf{1}\big\{\big|\langle \mathbf{A}, \mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_k} \rangle\big| = 2^{-s/2}\big\} \cdot A(i_1, \dots, i_k) \mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_k}$$

where we denote by $\mathbf{e}_{i_j}$ the canonical basis vectors in $\mathbb{R}^{d_j}$. Clearly, for all $\mathbf{u}_j \in \mathfrak{B}_{m_j, d_j}$,

$$\big\langle \mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle = \sum_{s=1}^{m_1 + \dots + m_k} \big\langle \mathbf{D}_s\big(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k\big), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle.$$

For a subset $\mathcal{T} \subset [d_1] \times \dots \times [d_k]$, the aspect ratio $\mu_{\mathcal{T}}$ is defined by

$$\mu_{\mathcal{T}} := \max_{\ell = 1, \dots, k} \max_{i_j : j \in [k] \setminus \ell} \mathrm{Card}\big(\big\{i_\ell : (i_1, \dots, i_k) \in \mathcal{T}\big\}\big).$$

Define the sampling locations

$$\Omega = \big\{(i_1, \dots, i_k) : \Delta(i_1, \dots, i_k) = 1\big\}$$

and the associated sampling operator

$$\mathcal{P}_\Omega(\mathbf{A}) = \sum_{i_1, \dots, i_k} \mathbf{1}\big((i_1, \dots, i_k) \in \Omega\big) A(i_1, \dots, i_k) \mathbf{e}_{i_1} \otimes \dots \otimes \mathbf{e}_{i_k}.$$

We shall now make use of the following version of the Chernoff bound:

Lemma 2: Let $X_1, \dots, X_n$ be independent binary random variables such that $\mathbb{P}(X_j = 1) = p_j \in [0, 1]$, $j = 1, \dots, n$. Then, for any $t \ge 0$,

$$\mathbb{P}\Big(\sum_{j=1}^n (X_j - p_j) \ge 2t \sqrt{\sum_{j=1}^n p_j(1 - p_j)}\Big) \le e^{-t^2}.$$

Lemma 2 is fairly standard and we include its proof in the Appendix for completeness.

By Lemma 2, there exists an absolute constant $C > 0$ such that for all $\alpha \ge 1$,

$$\mathbb{P}\Big(\mu_\Omega \ge C\alpha \max\Big\{\beta(\mathbf{P}), k \log d_{\max}\Big\}\Big) \le d_{\max}^{-\alpha}$$

where

$$\beta(\mathbf{P}) = \max_{j = 1, \dots, k} \max_{i_1, \dots, i_{j-1}, i_{j+1}, \dots, i_k} \sum_{i_j = 1}^{d_j} P(i_1, \dots, i_k)$$

and $d_{\max} := \max_{1 \le j \le k} d_j$. Denote the above event by $\mathcal{E}_1$ with $\mathbb{P}(\mathcal{E}_1) \ge 1 - d_{\max}^{-\alpha}$. The rest of our analysis is conditioned on event $\mathcal{E}_1$. Observe that

$$\big\langle \mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle = \sum_{s=1}^{m_1 + \dots + m_k} \big\langle \mathcal{P}_\Omega\big(\mathbf{D}_s(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k)\big), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle.$$

For $\mathbf{u}_j \in \mathfrak{B}_{m_j, d_j}$, let $\mathcal{A}_{b_j} = \big\{i_j : \big|u_j(i_j)\big| = 2^{-b_j/2}\big\}$ for $j = 1, \dots, k$. Then, we write

$$\mathbf{D}_s(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k) = \sum_{(b_1, \dots, b_k) : b_1 + \dots + b_k = s} \mathcal{P}_{\mathcal{A}_{b_1} \times \dots \times \mathcal{A}_{b_k}} \mathbf{D}_s\big(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k\big).$$

By definition of μ_Ω, on event $\mathcal{E}_1$, there exist $\tilde{\mathcal{A}}_{b_1} \subset \mathcal{A}_{b_1}, \dots, \tilde{\mathcal{A}}_{b_k} \subset \mathcal{A}_{b_k}$ such that

$$\big(\mathcal{A}_{b_1} \times \dots \times \mathcal{A}_{b_k}\big) \cap \Omega = \big(\tilde{\mathcal{A}}_{b_1} \otimes \dots \otimes \tilde{\mathcal{A}}_{b_k}\big) \cap \Omega$$

and

$$\mathrm{Card}^2(\tilde{\mathcal{A}}_{b_j}) \le \mu_\Omega \prod_{j=1}^k \mathrm{Card}(\tilde{\mathcal{A}}_{b_j}), \quad j = 1, 2, \dots, k.$$

We conclude with

$$\big\langle \mathbf{D}_s(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle = \sum_{s=1}^{m_1 + \dots + m_k} \sum_{b_1 + \dots + b_s = s} \big\langle \mathcal{P}_{\tilde{\mathcal{A}}_{b_1} \times \dots \times \tilde{\mathcal{A}}_{b_k}} \mathbf{D}_s(\mathbf{u}_1 \otimes \dots \otimes \mathbf{u}_k), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}} \big\rangle.$$

Given Ω, we define the balanced version of digitalization operator

$$\widetilde{\mathbf{D}}_s(\mathbf{u}_1 \otimes \ldots \otimes \mathbf{u}_k) = \sum_{(b_1,\ldots,b_k):b_1+\ldots+b_k=s} \mathcal{P}_{\tilde{\mathcal{A}}_{b_1}\times\ldots\times\tilde{\mathcal{A}}_{b_k}} \mathbf{D}_s(\mathbf{u}_1 \otimes \ldots \otimes \mathbf{u}_k)$$

where $\tilde{\mathcal{A}}_j$ are defined as above. Then, $\mathcal{P}_\Omega \mathbf{D}_s(\mathbf{u}_1 \otimes \ldots \mathbf{u}_k) = \mathcal{P}_\Omega \widetilde{\mathbf{D}}_s(\mathbf{u}_1 \otimes \ldots \mathbf{u}_k)$. Given Ω, define

$$\mathfrak{B}_{\Omega,m_\star} := \Big\{ \sum_{0\le s\le m_\star} \widetilde{\mathbf{D}}_s(\mathbf{u}_1 \otimes \ldots \mathbf{u}_k) + \sum_{m_\star < s \le m^\star} \mathbf{D}_s(\mathbf{u}_1 \otimes \ldots \otimes \mathbf{u}_k) : \mathbf{u}_j \in \mathfrak{B}_{m_j,d_j},\, j = 1,\ldots,k \Big\}$$

for any $0 < m_\star \le m^\star \le \sum_{j=1}^k m_j$. Conditioned on $\mathcal{E}_1$, we shall focus on $\{\Omega : \mu_\Omega \le \nu\}$ where $\nu = C\alpha \max\{\beta(\mathbf{P}), k\log d_{\max}\}$. Denote $\mathfrak{B}^\star_{\nu,m_\star} = \bigcup_{\mu_\Omega\le\nu} \mathfrak{B}^\star_{\Omega,m_\star}$. Following an identical argument as that in [20], we get

$$\big\|\boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\| \le 2^k \max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \langle \mathbf{Y}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\rangle.$$

The entropy number of $\mathfrak{B}^\star_{\nu,m_\star}$ plays an essential role in bounding $\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \langle \mathbf{Y}, \mathbf{X}\rangle$. Observe that $\mathfrak{B}^\star_{\nu,m_\star} \subset \mathfrak{B}_{m_1,d_1} \times \ldots \times \mathfrak{B}_{d_k,m_k}$ and

$$\begin{aligned}\mathrm{Card}\big(\mathfrak{B}_{m_j,d_j}\big) &\le \prod_{k=0}^{m_j} \binom{d_j}{2^k \wedge d_j} 2^{2^k \wedge d_j} \\ &\le \prod_{k=0}^{m_j} \exp\Big((2^k \wedge d_j)\big(\log 2 + 1 + (\log d_j/2^k)_+\big)\Big) \\ &\le \exp\Big(d_j \sum_{\ell=1}^{\infty} 2^{-\ell}\big(\log 2 + 1 + \log(2^\ell)\big)\Big) \\ &\le \exp\big(21 d_j/4\big),\end{aligned}$$

which implies that

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

See [20] for more details. More precise characterizations of $\mathrm{Card}(\mathfrak{B}^\star_{\nu,m_\star})$ can also be derived. For any $0 \le q \le s \le m_\star$, define

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

Lemma 3: Let $\nu \ge 1$. For all $0 \le q \le s \le m_\star$, the following bound holds

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

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

We write

$$\begin{aligned}\|\boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\| &\le 2^k \max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \big\langle \mathbf{Y}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle \\ &= 2^k \max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \Big(\sum_{0\le s\le m_\star} \big\langle \mathbf{D}_s(\mathbf{Y}), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle + \big\langle \mathbf{S}_\star(\mathbf{Y}), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle \Big)\end{aligned}$$

where $\mathbf{S}_\star(\mathbf{Y}) = \sum_{s>m_\star} \mathbf{D}_s(\mathbf{Y})$. The actual value of $m_\star$ is to be determined later.

Step 1: upper bound of $\big|\big\langle \mathbf{D}_s(\mathbf{Y}), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle\big|$. Recall the definition of $\mathfrak{D}_{\nu,s,q}$ and that

$$2^{-s} \le \|\mathbf{D}_s(\mathbf{Y})\|^2_{\ell_2} \le 1,$$

we can write

$$\mathbf{D}_s(\mathbf{Y}) \in \bigcup_{q=1}^{s} \big(\mathfrak{D}_{\nu,s,q} \setminus \mathfrak{D}_{\nu,s,q-1}\big).$$

Then

$$\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \big\langle \mathbf{D}_s(\mathbf{Y}), \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle = \max_{1\le q\le s} \max_{\mathbf{Y}_{s,q}\in\mathfrak{D}_{\nu,s,q}\setminus\mathfrak{D}_{\nu,s,q-1}} \big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle.$$

Observe that

$$\begin{aligned}&\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle \\ &= \sum_{i_j\in[d_j],j=1,\ldots,k} \frac{\Delta(i_1,\ldots,i_k)}{P(i_1\ldots i_k)} \varepsilon(i_1,\ldots,i_k) A(i_1,\ldots,i_k) \\ &\qquad Y_{s,q}(i_1,\ldots,i_k)\end{aligned}$$

where $\boldsymbol{\Delta}$ is a binary random tensor and $\boldsymbol{\varepsilon}$ is a Rademacher random tensor. Both of them have i.i.d. entries. By definition of $\mathbf{Y}_{s,q}$ and $\mathfrak{D}_{\nu,s,q}$, we have $\max_{i_1,\ldots,i_k} |Y_{s,q}(i_1,\ldots,i_k)| \le 2^{-s/2}$. Moreover,

$$\begin{aligned}&\mathrm{Var}\big(\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle\big) \\ &\quad = \sum_{i_j\in[d_j],j=1,\ldots,k} \frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} Y^2_{s,q}(i_1,\ldots,i_k).\end{aligned}$$

Since $\|\mathbf{Y}_{s,q}\|^2_{\mathrm{F}} \le 2^{q-s}$, we obtain

$$\begin{aligned}\mathrm{Var}\big(\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle\big) &\le \max_{i_j\in[d_j],j\in[k]} \frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} \|\mathbf{Y}_{s,q}\|^2_{\mathrm{F}} \\ &\le 2^{q-s} \max_{i_j\in[d_j],j\in[k]} \frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}.\end{aligned}$$

Recall the definition of $\alpha_\infty(\mathbf{A},\mathbf{P})$ and $\alpha_{2,\infty}(\mathbf{A},\mathbf{P})$. By Bernstein inequality for sum of bounded random variables, there exist absolute constants $C_0, C_1, C_2 > 0$ such that

$$\begin{aligned}&\mathbb{P}\Big(\big|\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle\big| \ge t\Big) \\ &\le \exp\Big(-\frac{C_0 t^2}{C_1 2^{q-s}\alpha^2_{2,\infty}(\mathbf{A},\mathbf{P}) + C_2 2^{-s/2} t\alpha_\infty(\mathbf{A},\mathbf{P})}\Big)\end{aligned}$$

for any $t > 0$. By the union bound and Lemma 3, we get

$$\begin{aligned}&\mathbb{P}\Big(\max_{\mathbf{Y}_{s,q}\in\mathfrak{D}_{\nu,s,q}} \big|\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon} \odot \widehat{\mathbf{A}}\big\rangle\big| \ge t\Big) \le \mathrm{Card}\big(\mathfrak{D}_{\nu,s,q}\big) \\ &\cdot \exp\Big(-\frac{C_0 t^2}{C_1 2^{q-s}\alpha^2_{2,\infty}(\mathbf{A},\mathbf{P}) + C_2 2^{-s/2} t\alpha_\infty(\mathbf{A},\mathbf{P})}\Big) \\ &\le \exp\Big(21\big(\sum_{j=1}^k d_j\big)/4 - \frac{C_0 t^2}{C_1 2^{q-s}\alpha^2_{2,\infty}(\mathbf{A},\mathbf{P})}\Big) \\ &+ \exp\Big(q s^k \log 2 + 2k^2 s^k \sqrt{\nu 2^q} L\big(\sqrt{\nu 2^q}, d_{\max} s^{k/2}\big) \\ &\qquad - \frac{C_0 2^{s/2} t^2}{C_2 t\alpha_\infty(\mathbf{A},\mathbf{P})}\Big).\end{aligned}$$

Recall that

$$0 \le q \le s \le m_\star \lesssim k \log d_{\max}$$

and

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

For large enough constants $C_3, C_4 > 0$, by choosing $t > 0$ such that

$$t \geq C_3 2^{(q-s)/2}\Big(\sum_{j=1}^k d_j\Big)^{1/2}\alpha_{2,\infty}(\mathbf{A},\mathbf{P}) + C_4 k^3 \log^{k+1} d_{\max}\sqrt{\nu}\, 2^{q-s}\alpha_\infty(\mathbf{A},\mathbf{P}),$$

we get for any $0 \leq q \leq s \leq m_\star$,

$$\mathbb{P}\Big(\max_{\mathbf{Y}_{s,q}\in\mathfrak{D}_{\nu,s,q}} \big|\big\langle \mathbf{Y}_{s,q}, \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big| \geq t\Big) \leq \exp\bigg(-\frac{C_0 t^2}{C_1 2^{q-s}\alpha_{2,\infty}^2(\mathbf{A},\mathbf{P})}\bigg) + \exp\bigg(-\frac{C_0 2^{s/2} t}{C_2\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg).$$

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

$$\mathbb{P}\Big(\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \Big|\sum_{0\leq s\leq m_\star}\big\langle \mathbf{D}_s(\mathbf{Y}), \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\Big| \geq (m_\star+1)t\Big) \leq \binom{m_\star+1}{2}\exp\bigg(-\frac{C_0 t^2}{C_1\alpha_{2,\infty}^2(\mathbf{A},\mathbf{P})}\bigg) + \binom{m_\star+1}{2}\exp\bigg(-\frac{C_0 t}{C_2\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg).$$

Step 2: upper bound of $\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}}\big|\big\langle \mathbf{S}_\star(\mathbf{Y}), \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big|$. For notation simplicity, we write $\mathbf{S}_\star$ in short for $\mathbf{S}_\star(\mathbf{Y})$. We apply Bernstein inequality to

$$\big\langle \mathbf{S}_\star, \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle = \sum_{i_j\in[d_j], j=1,\ldots,k} \frac{\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\varepsilon(i_1,\ldots,i_k)A(i_1,\ldots,i_k)S_\star(i_1,\ldots,i_k).$$

Clearly, $\big|S_\star(i_1,\ldots,i_k)\big| \leq 2^{-m_\star/2}$. Meanwhile,

$$\mathrm{Var}\big(\big\langle \mathbf{S}_\star, \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big) = \sum_{i_j\in[d_j], j=1,\ldots,k}\frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}S_\star^2(i_1,\ldots,i_k).$$

Following an identical approach as previously, we show that

$$\mathrm{Var}\big(\big\langle \mathbf{S}_\star, \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big) \leq \alpha_{2,\infty}^2(\mathbf{A},\mathbf{P}).$$

By Bernstein inequality and the union bound

$$\mathbb{P}\Big(\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \big|\big\langle \mathbf{S}_\star(\mathbf{Y}), \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big| \geq t\Big) \leq \mathrm{Card}\big(\mathfrak{B}^\star_{\nu,m_\star}\big)\exp\bigg(-\frac{C_0 t^2}{C_1\alpha_{2,\infty}^2(\mathbf{A},\mathbf{P}) + C_2 2^{-m_\star/2} t\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg) \leq \exp\bigg(21\sum_{j=1}^k d_j/4 - \frac{C_0 t^2}{C_1\alpha_{2,\infty}^2(\mathbf{A},\mathbf{P})}\bigg) + \exp\bigg(21\sum_{j=1}^k d_j/4 - \frac{C_0 2^{m_\star/2} t}{C_2\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg)$$

for some absolute constants $C_0, C_1, C_2 > 0$. For large enough constants $C_3, C_4 > 0$, by choosing t such that

$$t \geq C_3\Big(\sum_{j=1}^k d_j\Big)^{1/2}\alpha_{2,\infty}(\mathbf{A},\mathbf{P}) + C_4\Big(\sum_{j=1}^k d_j\Big)2^{-m_\star/2}\alpha_\infty(\mathbf{A},\mathbf{P}),$$

we obtain

$$\mathbb{P}\Big(\max_{\mathbf{Y}\in\mathfrak{B}^\star_{\nu,m_\star}} \big|\big\langle \mathbf{S}_\star(\mathbf{Y}), \boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\big\rangle\big| \geq t\Big) \leq \exp\bigg(-\frac{C_0 t^2}{C_1\alpha_{2,\infty}^2(\mathbf{A},\mathbf{P})}\bigg) + \exp\bigg(-\frac{C_0 2^{m_\star/2} t}{C_2\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg).$$

Step 3: finalize the proof of Theorem 2. Combining above bounds, we conclude that if for large enough constants $C_3, C_4, C_5 > 0$ such that

$$t \geq C_3\Big(\sum_{j=1}^k d_j\Big)^{1/2}\alpha_{2,\infty}(\mathbf{A},\mathbf{P}) + C_4 k^3\log^{k+1}(d_{\max})\sqrt{\nu}\alpha_\infty(\mathbf{A},\mathbf{P}) + C_5\Big(\sum_{j=1}^k d_j\Big)2^{-m_\star/2}\alpha_\infty(\mathbf{A},\mathbf{P}).$$

Thus

$$\mathbb{P}\Big(\|\boldsymbol{\varepsilon}\odot\widehat{\mathbf{A}}\| \geq (m_\star+2)t\Big) \leq \bigg(\binom{m_\star+1}{2}+1\bigg)\exp\bigg(-\frac{C_0 t^2}{C_1\alpha_2^2(\mathbf{A},\mathbf{P})}\bigg) + \bigg(\binom{m_\star+1}{2}+1\bigg)\exp\bigg(-\frac{C_0 t}{C_2\alpha_\infty(\mathbf{A},\mathbf{P})}\bigg).$$

Recall that $\nu = C_1\alpha\max\big\{\beta(\mathbf{P}), k\log d_{\max}\big\}$ and $m_\star \leq \sum_{j=1}^k 2\Big(\lceil\log_2\ d_j\rceil + 3\Big)$. By choosing $m_\star$ large enough such that $2^{-m_\star/2}\Big(\sum_{j=1}^k d_j\Big) \leq \sqrt{\nu}$, we conclude that for any $\gamma > 0$ such that

$$t \geq C_3\bigg(\Big(\sum_{j=1}^k d_j\Big)^{1/2} + \gamma k\log d_{\max}\bigg)\alpha_{2,\infty}(\mathbf{A},\mathbf{P}) + C_4\gamma k^3\log^{k+2}(d_{\max})\sqrt{\nu}\alpha_\infty(\mathbf{A},\mathbf{P}).$$

It follows immediately, by adjusting the constant C_3, that

$$\mathbb{P}\Big(\|\widehat{\mathbf{A}}-\mathbf{A}\| \geq t\Big) \leq 2d_{\max}^{-\gamma}.$$

C. Proof of Theorem 3

It suffices to prove the upper bound of $\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|$ where $\mathbf{M}_j = \mathcal{M}_j(\mathbf{A})$ and $\widehat{\mathbf{M}}_j = \mathcal{M}_j(\widehat{\mathbf{A}}^{\mathrm{SPA}})$. Without loss of generality, let $j = 1$. Recall the notation $d_{-1} = d_2\ldots d_k$. By denoting $\mathbf{E}_{i_1(i_2\ldots i_k)} \in \mathbb{R}^{d_1\times d_{-1}}$ the canonical basis matrices of $\mathbb{R}^{d_1\times d_{-1}}$ that is $\mathbf{E}_{i_1(i_2\ldots i_k)}$ has exactly value 1 on the $(i_1, i_2\ldots i_k)$ position and all 0's elsewhere. Then,

$$\widehat{\mathbf{M}}_j - \mathbf{M}_j = \sum_{i_j\in[d_j], 1\leq j\leq k}\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)\mathbf{E}_{i_1(i_2\ldots i_k)}$$

where $\mathbb{P}\big(\Delta(i_1,\ldots,i_k) = 1\big) = P(i_1,\ldots,i_k)$. We shall apply the matrix Bernstein inequality to bound the sum of random matrices for $\widehat{\mathbf{M}}_j - \mathbf{M}_j$. Denote the locations of small entries by

$$\boldsymbol{\Omega}_1 := \big\{(i_1,\ldots,i_k) : \|A(i_1,\ldots,i_k)\| \leq \|\mathbf{A}\|_{\mathrm{F}}/(d_1\ldots d_k)^{1/2}\big\} \subset [d_1]\times\ldots\times[d_k]$$

moderate entries by

$$\Omega_2 := \big\{(i_1,\ldots,i_k) : \|A(i_1,\ldots,i_k)\|/\|\mathbf{A}\|_{\mathrm{F}} \in \big(1/(d_1\ldots d_k)^{1/2}, 1/n^{1/2}\big)\big\} \subset [d_1]\times\ldots\times[d_k]$$

and large entries by

$$\Omega_3 := \big\{(i_1,\ldots,i_k) : \|A(i_1,\ldots,i_k)\| \geq \|\mathbf{A}\|_{\mathrm{F}}/n^{1/2}\big\} \subset [d_1]\times\ldots\times[d_k].$$

Recall that $P(i_1,\ldots,i_k)=1$ for $(i_1,\ldots,i_k)\in\Omega_3$. Then, for any $(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2$, we have

$$\Big\|\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)\mathbf{E}_{i_1(i_2\ldots i_k)}\Big\| \leq \max_{i_j\in[d_j],1\leq j\leq k}\Big|\frac{A(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\Big|.$$

Moreover,

$$\begin{aligned}&\Big\|\sum_{i_j\in[d_j],1\leq j\leq k}\mathbb{E}\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)^2 \cdot \mathbf{E}_{i_1(i_2\ldots i_k)}\mathbf{E}_{i_1(i_2\ldots i_k)}^{\top}\Big\| \\ &\leq \max_{1\leq i_1\leq d_1}\sum_{i_j\in[d_j],2\leq j\leq k}\frac{A^2(i_1,\ldots,i_k)\big(1-P(i_1,\ldots,i_k)\big)}{P(i_1,\ldots,i_k)} \\ &\leq \max_{1\leq i_1\leq d_1}\sum_{i_j\in[d_j],j\geq 2,(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2}\frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}.\end{aligned}$$

Similarly,

$$\begin{aligned}&\Big\|\sum_{i_j\in[d_j],1\leq j\leq k}\mathbb{E}\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)^2 \cdot \mathbf{E}_{i_1(i_2\ldots i_k)}^{\top}\mathbf{E}_{i_1(i_2\ldots i_k)}\Big\| \\ &\leq \max_{i_j\in[d_j],2\leq j\leq k}\sum_{i_1=1}^{d_1}\frac{A^2(i_1,\ldots,i_k)\big(1-P(i_1,\ldots,i_k)\big)}{P(i_1,\ldots,i_k)} \\ &\leq \max_{i_j\in[d_j],2\leq j\leq k}\sum_{i_1\in[d_1],(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2}\frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}.\end{aligned}$$

Observe that if $(i_1,\ldots,i_k)\in\Omega_1$, then

$$\Big|\frac{A(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\Big| = \frac{(d_1\ldots d_k)}{n}|A(i_1,\ldots,i_k)| \leq \frac{(d_1\ldots d_k)^{1/2}}{n}\|\mathbf{A}\|_{\mathrm{F}}$$

and

$$\frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} = \frac{(d_1\ldots d_k)A^2(i_1,\ldots,i_k)}{n} \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}.$$

Similarly, if $(i_1,\ldots,i_k)\in\Omega_2$, then

$$\Big|\frac{A(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\Big| = \frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n|A(i_1,\ldots,i_k)|} \leq \frac{(d_1\ldots d_k)^{1/2}}{n}\|\mathbf{A}\|_{\mathrm{F}}$$

and

$$\frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} = \frac{\|\mathbf{A}\|_{\mathrm{F}}^2}{n}.$$

By matrix Bernstein inequality [32], for any $t\geq 0$, with probability at least $1-e^{-t}$ that

$$\big\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\big\| \leq 2\|\mathbf{A}\|_{\mathrm{F}}\bigg(\sqrt{\frac{d_2d_3\ldots d_k(t+k\log d_{\max})}{n}} + \frac{(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})}{n}\bigg).$$

Since $\widehat{\mathbf{M}}_j = \mathbf{M}_j + \big(\widehat{\mathbf{M}}_j - \mathbf{M}_j\big)$, the claim follows directly from Davis-Kahan Thoerem as in (3).

D. Proof of Theorem 4

Theorem 4 is an immediate consequence of the following concentration bound.

Lemma 4: Let $\mathbf{U}_j^{(r_j)}$ be the r_j leading left singular vectors of $\mathcal{M}_j(\mathbf{A})$, and $\widehat{\mathbf{U}}_j^{(r_j)}$ be the output from Algorithm 2. There exist absolute constants $C_1, C_2 > 0$ such that if

$$n \geq C_1(d_1\ldots d_k)^{1/2}(t+k\log d_{\max}),$$

then for any $t\geq 0$, the following bound holds with probability at least $1-e^{-t}$:

$$\begin{aligned}&\big\|\widehat{\mathbf{U}}_j^{(r_j)}\big(\widehat{\mathbf{U}}_j^{(r_j)}\big)^{\top} - \mathbf{U}_j^{(r_j)}\big(\mathbf{U}_j^{(r_j)}\big)^{\top}\big\| \\ &\leq C_2\frac{\|\mathbf{A}\|_{\mathrm{F}}}{\bar{g}_{r_j}(\mathbf{M}_j\mathbf{M}_j^{\top})}\bigg(\sigma_{\max}(\mathbf{M}_j)\sqrt{\frac{d_j(t+k\log d_{\max})}{n}} + \|\mathbf{A}\|_{\mathrm{F}}\frac{(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})}{n}\bigg).\end{aligned}$$

Proof of Lemma 4: With out loss of generality, we assume $j=1$ without loss of generality. In this case, $\widehat{\mathbf{M}}_j^{(1)} = \mathcal{M}_j(\widehat{\mathbf{A}}_1^{\mathrm{SPA}})$, $\widehat{\mathbf{M}}_j^{(2)} = \mathcal{M}_j(\widehat{\mathbf{A}}_2^{\mathrm{SPA}}) \in \mathbb{R}^{d_1\times(d_2\ldots d_k)}$. Observe that

$$\begin{aligned}\widehat{\mathbf{M}}_j^{(1)}\big(\widehat{\mathbf{M}}_j^{(2)}\big)^{\top} &= \mathbf{M}_j\mathbf{M}_j^{\top} + \big(\widehat{\mathbf{M}}_j^{(1)} - \mathbf{M}_j\big)\mathbf{M}_j^{\top} \\ &\quad + \mathbf{M}_j\big(\widehat{\mathbf{M}}_j^{(2)} - \mathbf{M}_j\big)^{\top} + \big(\widehat{\mathbf{M}}_j^{(1)} - \mathbf{M}_j\big)\big(\widehat{\mathbf{M}}_j^{(2)} - \mathbf{M}_j\big)^{\top}.\end{aligned}$$

Step 1: upper bound of $\big\|\big(\widehat{\mathbf{M}}_j^{(1)} - \mathbf{M}_j\big)\big(\widehat{\mathbf{M}}_j^{(2)} - \mathbf{M}_j\big)^{\top}\big\|$. Denote by $\mathbf{Z}_1 = \widehat{\mathbf{M}}_j^{(1)} - \mathbf{M}_j$. By Theorem 3, there exists an event $\mathcal{E}_1$ with $\mathbb{P}(\mathcal{E}_1)\geq 1-e^{-t}$ such that on event $\mathcal{E}_1$,

$$\|\mathbf{Z}_1\| \leq C\|\mathbf{A}\|_{\mathrm{F}}\bigg(\sqrt{\frac{d_2d_3\ldots d_k(t+k\log d_{\max})}{n}} + \frac{(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})}{n}\bigg).$$

Denote by $\|\mathbf{Z}_1\|_{2,\infty}$ the maximal column ℓ_2 norm., i.e., $\|\mathbf{Z}_1\|_{2,\infty} = \max_{j\in[d_2\ldots d_k]}\big\|\mathbf{Z}_1\mathbf{e}_j\big\|_{\ell_2}$. Clearly, there exists

an absolute constant $C_1 > 0$ such that

$$\begin{aligned}
&\|\mathbf{Z}_1\|_{2,\infty} \\
\le& C_1\Bigg(\max_{i_j\in[d_j],2\le j\le k}\sqrt{\sum_{i_1\in[d_1]:(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\frac{A^2(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}} \\
&\qquad\cdot\sqrt{t+k\log d_{\max}} \\
&\quad+\max_{(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\left|\frac{A(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}\right|(t+k\log d_{\max})\Bigg) \\
\le& C_1\|\mathbf{A}\|_{\mathrm{F}}\Bigg(\sqrt{\frac{d_1(t+k\log d_{\max})}{n}} \\
&\quad+\frac{(d_1\dots d_k)^{1/2}(t+k\log d_{\max})}{n}\Bigg),
\end{aligned}$$

which holds with probability at least $1-e^{-t}$. Denote the above event by $\mathcal{E}_3$. We shall proceed conditional on $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$. Write

$$\begin{aligned}
\mathbf{Z}_1\big(\widehat{\mathbf{M}}_j^{(2)}-\mathbf{M}_j\big)^\top=\sum_{i_j\in[d_j],1\le j\le k}\Big(&\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)} \\
&-A(i_1,\dots,i_k)\Big)\mathbf{Z}_1\mathbf{E}_{i_1(i_2\dots i_k)}^\top
\end{aligned}$$

which is again a sum of random matrices. Clear, for any $(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2$,

$$\begin{aligned}
&\left\|\Big(\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}-A(i_1,\dots,i_k)\Big)\mathbf{Z}_1\mathbf{E}_{i_1(i_2\dots i_k)}^\top\right\| \\
&\quad\le\max_{(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\left|\frac{A(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}\right|\|\mathbf{Z}_1\|_{2,\infty} \\
&\quad\le\frac{(d_1\dots d_k)^{1/2}}{n}\|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{Z}_1\|_{2,\infty}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\Bigg\|\sum_{i_j\in[d_j],1\le j\le k}\mathbb{E}\Big(\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}-A(i_1,\dots,i_k)\Big)^2 \\
&\qquad\cdot\mathbf{Z}_1\mathbf{E}_{i_1(i_2\dots i_k)}^\top\mathbf{E}_{i_1(i_2\dots i_k)}\mathbf{Z}_1^\top\Bigg\| \\
&\le\max_{i_j\in[d_j],2\le j\le k}\|\mathbf{Z}_1\|^2\sum_{i_1\in[d_1]:(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\frac{A^2(i_1,\dots,i_k)}{P(i_1,\dots,i_k)} \\
&\le\frac{d_1\|\mathbf{A}\|_{\mathrm{F}}^2}{n}\|\mathbf{Z}_1\|^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Bigg\|\sum_{i_j\in[d_j],1\le j\le k}\mathbb{E}\Big(\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}-A(i_1,\dots,i_k)\Big)^2 \\
&\qquad\mathbf{E}_{i_1(i_2\dots i_k)}\mathbf{Z}_1^\top\mathbf{Z}_1\mathbf{E}_{i_1(i_2\dots i_k)}^\top\Bigg\| \\
&\le\max_{(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\frac{A^2(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}\|\mathbf{Z}_1\|_{\mathrm{F}}^2 \\
&\le\frac{d_1\|\mathbf{A}\|_{\mathrm{F}}^2}{n}\|\mathbf{Z}_1\|^2.
\end{aligned}$$

By matrix Bernstein inequality, the following bound holds with probability at least $1-e^{-t}$,

$$\begin{aligned}
&\big\|\big(\widehat{\mathbf{M}}_j^{(1)}-\mathbf{M}_j\big)\big(\widehat{\mathbf{M}}_j^{(2)}-\mathbf{M}_j\big)^\top\big\| \\
&\quad\le C\|\mathbf{A}\|_{\mathrm{F}}\Bigg(\sqrt{\frac{d_1(t+k\log d_{\max})}{n}}\|\mathbf{Z}_1\| \\
&\qquad+\frac{(d_1\dots d_k)^{1/2}(t+k\log d_{\max})}{n}\|\mathbf{Z}_1\|_{2,\infty}\Bigg).
\end{aligned}$$

Denote the above event by $\mathcal{E}_4$. On event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$, if

$$n\ge C_1(d_1d_2\dots d_k)^{1/2}(t+k\log d_{\max}),$$

then

$$\begin{aligned}
&\big\|\big(\widehat{\mathbf{M}}_j^{(1)}-\mathbf{M}_j\big)\big(\widehat{\mathbf{M}}_j^{(2)}-\mathbf{M}_j\big)^\top\big\| \\
&\quad\le C_2\|\mathbf{A}\|_{\mathrm{F}}^2\Bigg(\frac{(d_1\dots d_k)^{1/2}(t+k\log d_{\max})}{n} \\
&\qquad+\frac{d_1^{1/2}(d_1\dots d_k)^{1/2}(t+k\log d_{\max})^{3/2}}{n^{3/2}}\Bigg) \\
&\quad\le C_2\|\mathbf{A}\|_{\mathrm{F}}^2\frac{(d_1\dots d_k)^{1/2}(t+k\log d_{\max})}{n}.
\end{aligned}$$

Step 2: upper bound of $\big\|\mathbf{M}_j\big(\widehat{\mathbf{M}}_j^{(2)}-\mathbf{M}_j\big)^\top\big\|$. We write

$$\begin{aligned}
\mathbf{M}_j\big(\widehat{\mathbf{M}}_j^{(2)}-\mathbf{M}_j\big)^\top=\sum_{i_j\in[d_j],1\le j\le k}\Big(&\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)} \\
&-A(i_1,\dots,i_k)\Big)\mathbf{M}_j\mathbf{E}_{i_1(i_2\dots i_k)}^\top.
\end{aligned}$$

The proof follows identically as above. Indeed, for any $(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2$,

$$\begin{aligned}
&\left\|\left(\frac{A(i_1,\dots,i_k)\Delta(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}-A(i_1,\dots,i_k)\right)\right. \\
&\qquad\left.\cdot\mathbf{M}_j\mathbf{E}_{i_1(i_2\dots i_k)}^\top\right\| \\
&\le\max_{(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}\left|\frac{A(i_1,\dots,i_k)}{P(i_1,\dots,i_k)}\right| \\
&\qquad\cdot\max_{i_j\in[d_j],2\le j\le k}\sqrt{\sum_{i_1:(i_1,\dots,i_k)\in\Omega_1\cup\Omega_2}A^2(i_1,\dots,i_k)} \\
&\le\frac{(d_1\dots d_k)^{1/2}}{n}\Big(\frac{d_1}{n}\Big)^{1/2}\|\mathbf{A}\|_{\mathrm{F}}^2.
\end{aligned}$$

Moreover,

$$\Big\| \sum_{i_j\in[d_j],1\le j\le k} \mathbb{E}\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)^2 \cdot \mathbf{M}_j\mathbf{E}^\top_{i_1(i_2\ldots i_k)}\mathbf{E}_{i_1(i_2\ldots i_k)}\mathbf{M}_j^\top \Big\|$$

$$\le \max_{i_j\in[d_j],2\le j\le k} \sum_{i_1:(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2} \frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\|\mathbf{M}_j\|^2$$

$$\le \frac{d_1}{n}\|\mathbf{A}\|_{\mathrm{F}}^2\sigma^2_{\max}(\mathbf{M}_j).$$

Similarly,

$$\Big\| \sum_{i_j\in[d_j],1\le j\le k} \mathbb{E}\Big(\frac{A(i_1,\ldots,i_k)\Delta(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)} - A(i_1,\ldots,i_k)\Big)^2 \cdot \mathbf{E}_{i_1(i_2\ldots i_k)}\mathbf{M}_j^\top\mathbf{M}_j\mathbf{E}^\top_{i_1(i_2\ldots i_k)} \Big\|$$

$$\le \Big(\max_{i_j\in[d_j],2\le j\le k} \sum_{i_1:(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2} A^2(i_1,\ldots,i_k)\Big) \cdot \Big(\max_{i_1\in[d_1]} \sum_{i_j\in[d_j],2\le j\le k:(i_1,\ldots,i_k)\in\Omega_1\cup\Omega_2} \frac{A^2(i_1,\ldots,i_k)}{P(i_1,\ldots,i_k)}\Big)$$

$$\le \frac{d_1d_2\ldots d_k}{n^2}\|\mathbf{A}\|_{\mathrm{F}}^4.$$

By matrix Bernstein inequality [32], if $n \ge C_1(d_1\ldots d_k)^{1/2}(t + k\log d_{\max})$, then with probability at least $1-e^{-t}$ such that

$$\big\|\mathbf{M}_j\big(\widehat{\mathbf{M}}_j^{(2)} - \mathbf{M}_j\big)^\top\big\| \le C_2\|\mathbf{A}\|_{\mathrm{F}}\Big(\sigma_{\max}(\mathbf{M}_j)\sqrt{\frac{d_1(t+k\log d_{\max})}{n}} + \|\mathbf{A}\|_{\mathrm{F}}\frac{(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})}{n}\Big).$$

Denote this event by $\mathcal{E}_5$. Clearly, an identical bound holds for $\big\|\big(\widehat{\mathbf{M}}_j^{(1)} - \mathbf{M}_j\big)\mathbf{M}_j^\top\big\|$ with the same probability. Denote this event by $\mathcal{E}_6$.

Final step: finalize the proof of Theorem 4. On event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4\cap\mathcal{E}_5\cap\mathcal{E}_6$, if $n \ge C_1(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})$, there exists an absolute constant $C_2>0$ such that

$$\big\|\widehat{\mathbf{M}}_j^{(1)}\big(\widehat{\mathbf{M}}_j^{(2)}\big)^\top - \mathbf{M}_j\mathbf{M}_j^\top\big\| \le C_2\|\mathbf{A}\|_{\mathrm{F}}\Big(\sigma_{\max}(\mathbf{M}_j)\sqrt{\frac{d_1(t+k\log d_{\max})}{n}} + \|\mathbf{A}\|_{\mathrm{F}}\frac{(d_1\ldots d_k)^{1/2}(t+k\log d_{\max})}{n}\Big),$$

which concludes the proof by adjusting the constant C_2 and applying Davis-Kahan Theorem. □

APPENDIX

A. Proof of Lemma 2

Clearly, for any t and $\lambda>0$,

$$\begin{aligned}\mathbb{P}\Big(\sum_{j=1}^n(X_j-p_j)\ge t\Big) &= \mathbb{P}\Big(\exp\Big\{\lambda\sum_{j=1}^n(X_j-p_j)\Big\}\ge\exp\{\lambda t\}\Big)\\ &\le e^{-\lambda t}\mathbb{E}\exp\Big\{\lambda\sum_{j=1}^n(X_j-p_j)\Big\}\\ &\le e^{-\lambda t}\prod_{j=1}^n\mathbb{E}e^{\lambda(X_j-p_j)}\\ &\le e^{-\lambda t}\prod_{j=1}^n\big(p_je^{\lambda(1-p_j)}+(1-p_j)e^{-\lambda p_j}\big).\end{aligned}$$

Note that $e^x\le 1+x+x^2$ for any $x\in[-1,1]$. Then,

$$p_je^{\lambda(1-p_j)}+(1-p_j)e^{-\lambda p_j} \le 1+\lambda^2p_j(1-p_j)\le e^{\lambda^2p_j(1-p_j)}.$$

Therefore, we obtain

$$\mathbb{P}\Big(\sum_{j=1}^n(X_j-p_j)\ge t\Big)\le e^{-\lambda t}\prod_{j=1}^n e^{\lambda^2p_j(1-p_j)} = \exp\Big\{-\lambda t+\lambda^2\sum_{j=1}^n p_j(1-p_j)\Big\}.$$

By choosing $\lambda = t/2\sum_{j=1}^n p_j(1-p_j)$, we end up with

$$\mathbb{P}\Big(\sum_{j=1}^n(X_j-p_j)\ge t\Big)\le\exp\Big\{-t^2/4\sum_{j=1}^n p_j(1-p_j)\Big\}.$$

The proof is closed after choosing $t = 2s\sqrt{\sum_{j=1}^n p_j(1-p_j)}$ for $s\ge 0$.

B. Proof of Lemma 3

The proof follows from the same argument as that for Lemma 12 of [20]. More specifically, denote the aspect ratio for a block $A_1\times\ldots A_k\subset[d_1]\times\ldots\times[d_k]$,

$$h(A_1\times\ldots\times A_k) = \min\Big\{\nu : |A_j|^2\le\nu\prod_{j=1}^k|A_j|,\ j=1,2,\ldots,k\Big\}.$$

We bound the entropy of a single block. Let

$$\mathfrak{D}^{(\mathrm{block})}_{\nu,\ell} = \Big\{\mathrm{sgn}(u_1(a_1))\ldots\mathrm{sgn}(u_k(a_k))\mathbf{1}\{(a_1,\ldots,a_k)\in A_1\times\ldots\times A_k\}: h(A_1\times\ldots A_k)\le\nu,\ \prod_{j=1}^k|A_j|=\ell\Big\}.$$

By definition, we obtain

$$\max\left(|A_1|^2,\ldots,|A_k|^2\right)\le \nu|A_1||A_2|\ldots|A_k|\le \nu\ell.$$

By dividing $\mathfrak{D}_{\nu,\ell}^{(\text{block})}$ into subsets according to $(\ell_1,\ldots,\ell_k)=(|A_1|,\ldots,|A_k|)$, we find

$$\left|\mathfrak{D}_{\nu,\ell}^{(\text{block})}\right|\le \sum_{\ell_1\ldots\ell_k=\ell,\max_j \ell_j\le\sqrt{\nu\ell}} 2^{\ell_1+\ldots+\ell_k}\binom{d_1}{\ell_1}\ldots\binom{d_k}{\ell_k}.$$

By the Stirling formula, for $j=1,2,\ldots,k$,

$$\binom{d_j}{\ell_j}\le\frac{d_j^{\ell_j}}{(\ell_j!)}\le\left(\frac{d_j}{\ell_j}\right)^{\ell_j}e^{\ell_j}\frac{1}{\sqrt{2\pi\ell_j}},$$

then

$$\log\left[\sqrt{2\pi\ell_j}2^{\ell_j}\binom{d_j}{\ell_j}\right]\le\ell_j L(\ell_j,2d_{\max})\le\sqrt{\nu\ell}L(\sqrt{\nu\ell},2d_{\max})$$

where $L(x,y):=\max\{1,\log(ey/x)\}$. Let $\ell=\prod_{j=1}^m p_j^{v_j}$ with distinct prime factors p_j. Since $(v_j+1)v_j/(2p_j^{v_j/2})$ is upper bounded by 2.66 for $p_j=2$, by 1.16 for $p_j=3$ and by 1 for $p_j\ge 5$, we get

$$\begin{aligned}\left|\{(\ell_1,\ldots,\ell_k):\ell_1\ldots\ell_k=\ell\}\right|&=\prod_{j=1}^m\binom{v_j+1}{k-1}\\&\le\prod_{j=1}^m\binom{v_j+1}{2}^{k/2}\\&\le(2.66\times1.16)^{k/2}(\sqrt{\ell})^{k/2}\\&\le\prod_{j=1}^k\left(2\sqrt{2\pi\ell_j}\right)^{k/2},\qquad\forall\prod_{j=1}^k\ell_j=\ell.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left|\mathfrak{D}_{\nu,\ell}^{(\text{block})}\right|\\&\le\frac{\exp\left(k\sqrt{\nu\ell}L(\sqrt{\nu\ell},2d_{\max})\right)}{\prod_{j=1}^k\sqrt{2\pi\ell_j}}\prod_{j=1}^k\left(2\sqrt{2\pi\ell_j}\right)^{k/2},\\&\qquad\qquad\forall(\ell_1\ldots\ell_k)=\ell\\&\le2^{k^2/2}(2\pi)^{k(k-2)/4}\ell^{(k-2)/4}\exp\left(k\sqrt{\nu\ell}L(\sqrt{\nu\ell},2d_{\max})\right)\\&\le2^{k^2/2}(2\pi)^{k(k-2)/4}\exp\left(2k\sqrt{\nu\ell}L(\sqrt{\nu\ell},2d_{\max})\right).\end{aligned}$$

Due to the constraint $b_1+b_2+\ldots+b_k=s$ in defining $\mathfrak{B}_{\nu,m_\star}^\star$, for any $\mathbf{Y}\in\mathfrak{B}_{\nu,m_\star}^\star$, $\mathbf{D}_s(\mathbf{Y})$ is composed of at most $i^\star:=\binom{s+k-1}{k-1}$ blocks. Since the sum of the sizes of the blocks is bounded by 2^q, we obtain

$$\begin{aligned}\left|\mathfrak{D}_{\nu,s,q}\right|&\le\sum_{\ell_1+\ldots+\ell_{i^\star}\le2^q}\prod_{i=1}^{i^\star}\left|\mathfrak{D}_{\nu,\ell_i}^{(\text{block})}\right|\\&\le\sum_{\ell_1+\ldots+\ell_{i^\star}\le2^q}(2\pi)^{i^\star k(k-2)/4}2^{i^\star k^2/2}\\&\qquad\cdot\exp\left(2k\sum_{i=1}^{i^\star}\sqrt{\nu\ell_i}L(\sqrt{\nu\ell_i},2d_{\max})\right)\\&\le2^{i^\star k^2/2}(2^q)^{i^\star}(2\pi)^{i^\star k(k-2)/4}\\&\qquad\cdot\max_{\ell_1+\ldots+\ell_{i^\star}\le2^q}\exp\left(2k\sum_{i=1}^{i^\star}\sqrt{\nu\ell_i}L(\sqrt{\nu\ell_i},2d_{\max})\right).\end{aligned}$$

As shown in [20], $\sum_{i=1}^{i^\star}\sqrt{\ell_i}L(\sqrt{\nu\ell_i},2d_{\max})\le\sqrt{i^\star 2^q}\left(L(\sqrt{\nu 2^q},2d_{\max})+\log(\sqrt{i^\star})\right)$, we obtain

$$\begin{aligned}\log\left|\mathfrak{D}_{\nu,s,q}\right|&\le i^\star\log(2^q)+i^\star k(k-2)/2+i^\star k^2/2\\&\qquad+2k\sqrt{i^\star\nu 2^q}L\left(\sqrt{\nu 2^q},2d_{\max}\sqrt{i^\star}\right).\end{aligned}$$

Since $i^\star=\binom{s+k-1}{k-1}\le s^k$, it follows that

$$\log\left|\mathfrak{D}_{\nu,s,q}\right|\le qs^k\log 2+2k^2s^k\sqrt{\nu 2^q}L\left(\sqrt{\nu 2^q},d_{\max}s^{k/2}\right).$$

REFERENCES

[1] Y. Kluger, "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Res.*, vol. 13, no. 4, pp. 703–716, Apr. 2003.

[2] U. Stelzl *et al.*, "A human protein-protein interaction network: A resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, Sep. 2005.

[3] S. M. Smith *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, Jan. 2004.

[4] R. Orús, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," *Ann. Phys.*, vol. 349, pp. 117–158, Oct. 2014.

[5] 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.

[6] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 517–520.

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

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.

[9] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[10] J. Scott, *Social Network Analysis*. Newbury Park, CA, USA: Sage, 2017.

[11] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Found. Trend Theor. Comput. Sci.*, vol. 10, no. 2, pp. 1–157, 2014.

[12] A. Frieze, R. Kannan, and S. Vempala, "Fast monte-carlo algorithms for finding low-rank approximations," *J. ACM*, vol. 51, no. 6, pp. 1025–1041, Nov. 2004.

[13] S. Arora, E. Hazan, and S. Kale, "A fast random sampling algorithm for sparsifying matrices," in in *Proc. APPROX-RANDOM*, vol. 6. Cham, Switzerland: Springer, 2006, pp. 272–279.

[14] D. Achlioptas and F. Mcsherry, "Fast computation of low-rank matrix approximations," *J. ACM*, vol. 54, no. 2, p. 9, Apr. 2007.

[15] P. Drineas and A. Zouzias, "A note on element-wise matrix sparsification via a matrix-valued bernstein inequality," *Inf. Process. Lett.*, vol. 111, no. 8, pp. 385–389, Mar. 2011.
[16] D. Achlioptas, Z. S. Karnin, and E. Liberty, "Near-optimal entrywise sampling for data matrices," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1565–1573.
[17] A. Krishnamurthy and A. Singh, "Low-rank matrix and tensor completion via adaptive sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 836–844.
[18] 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, Sep. 2015.
[19] S. Bhojanapalli and S. Sanghavi, "A new sampling technique for tensors," 2015, *arXiv:1502.05023*. [Online]. Available: http://arxiv.org/abs/1502.05023
[20] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Found. Comput. Math.*, vol. 16, no. 4, pp. 1031–1068, Aug. 2016.
[21] M. Yuan and C.-H. Zhang, "Incoherent tensor norms and their applications in higher order tensor completion," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6753–6766, Oct. 2017.
[22] D. Xia and M. Yuan, "On polynomial time methods for exact low rank tensor completion," 2017, *arXiv:1702.06980*. [Online]. Available: http://arxiv.org/abs/1702.06980
[23] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.
[24] 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, Jul. 2017.
[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3. Baltimore, MD, USA: JHU Press, 2012.
[26] M. W. Berry, "Large-scale sparse singular value computations," *Int. J. Supercomputing Appl.*, vol. 6, no. 1, pp. 13–49, Apr. 1992.
[27] M. Kobayashi, G. Dupret, O. King, and H. Samukawa, "Estimation of singular values of very large matrices using random sampling," *Comput. Math. Appl.*, vol. 42, nos. 10–11, pp. 1331–1352, Nov. 2001.
[28] M. Holmes, A. Gray, and C. Isbell, "Fast SVD for large-scale matrices," in *Proc. Workshop Efficient Mach. Learn. (NIPS)*, vol. 58, 2007, pp. 249–252.
[29] A. K. Menon and C. Elkan, "Fast algorithms for approximating the singular value decomposition," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 2, p. 13, 2011.
[30] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numer. Anal.*, vol. 7, no. 1, pp. 1–46, Mar. 1970.
[31] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Subspace sampling and relative-error matrix approximation: Column-based methods," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Cham, Switzerland: Springer, 2006, pp. 316–326.
[32] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, Aug. 2012.