LOWER MEMORY OBLIVIOUS (TENSOR) SUBSPACE EMBEDDINGS WITH FEWER RANDOM BITS: MODEWISE METHODS FOR LEAST SQUARES[*]

MARK A. IWEN[†], DEANNA NEEDELL[‡], ELIZAVETA REBROVA[‡], AND ALI ZARE[§]

Abstract. In this paper new general modewise Johnson–Lindenstrauss (JL) subspace embeddings are proposed that can be both generated much faster and stored more easily than traditional JL embeddings when working with extremely large vectors and/or tensors. Corresponding embedding results are then proven for two different types of low-dimensional (tensor) subspaces. The first of these new subspace embedding results produces improved space complexity bounds for embeddings of rank-r tensors whose CP decompositions are contained in the span of a fixed (but unknown) set of r rank-1 basis tensors. In the traditional vector setting this first result yields new and very general near-optimal oblivious subspace embedding constructions that require fewer random bits to generate than standard JL embeddings when embedding subspaces of $\mathbb{C}^N$ spanned by basis vectors with special Kronecker structure. The second result proven herein provides new fast JL embeddings of arbitrary r-dimensional subspaces $\mathcal{S} \subset \mathbb{C}^N$ which also require fewer random bits (and so are easier to store, i.e., require less space) than standard fast JL embedding methods in order to achieve small ε-distortions. These new oblivious subspace embedding results work by (i) effectively folding any given vector in $\mathcal{S}$ into a (not necessarily low-rank) tensor, and then (ii) embedding the resulting tensor into $\mathbb{C}^m$ for $m \leq Cr \log^c(N)/\varepsilon^2$. Applications related to compression and fast compressed least squares solution methods are also considered, including those used for fitting low-rank CP decompositions, and the proposed JL embedding results are shown to work well numerically in both settings.

Key words. Johnson–Lindenstrauss embeddings, low-rank tensors, tensors, least squares fitting, CP decompositions, dimensionality reduction, fast approximation algorithms

AMS subject classifications. 15A69, 15B52, 15-04, 65F30, 68Q87

DOI. 10.1137/19M1308116

1. Motivation and applications. Due to the recent explosion of massively large-scale data, the need for geometry preserving dimension reduction has become important in a wide array of applications in signal processing (see, e.g., [23, 22, 4, 62, 26, 14]) and data science (see, e.g., [8, 15]). This reduction is possible even on large-dimensional objects when the class of such objects possesses some sort of lower-dimensional intrinsic structure. For example, in classical compressed sensing [23, 22] and its related streaming applications [18, 19, 25, 31], the signals of interest are *sparse* vectors—vectors whose entries are mostly zero. In matrix recovery [15, 49], one often analogously assumes that the underlying matrix is low-rank. Under such models, tools such as the Johnson–Lindenstrauss lemma [33, 2, 20, 37, 38] and the related restricted isometry property [16, 7] ask that the geometry of the signals be

[†]Department of Mathematics, Michigan State University, East Lansing, MI 48824 USA (iwenmark@msu.edu).
[‡]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095 USA (deanna@math.ucla.edu, rebrova@math.ucla.edu).
[§]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824 USA (zareali@msu.edu).

preserved after projection into a lower-dimensional space. Typically, such projections are obtained via random linear maps that map into a dimension much smaller than the ambient dimension of the domain; s-sparse n-dimensional vectors can be projected into a dimension that scales like $s \log(n)$, and $n \times n$ rank-r matrices can be recovered from $O(rn)$ linear measurements [23, 22, 15]. Then, inference tasks or reconstruction can be performed from those lower-dimensional representations.

Here, our focus is on dimension reduction of *tensors*, multiway arrays that appear in an abundance of large-scale applications, including video and longitudinal imaging [40, 11], machine learning [50, 57], and differential equations [10, 41]. Although tensors are a natural extension of matrices, their complicated structure leads to challenges both in defining low-dimensional structure and in dimension reduction projections. In particular, there are many notions of tensor rank, and various techniques exist for computing the corresponding decompositions [36, 61]. In this paper, we focus on tensors with low CP-rank, tensors that can be written as a sum of a few rank-1 tensors written as outer products of basis vectors. The CP-rank and CP-decompositions are natural extensions of matrix rank and SVD and are well motivated by applications in topic modeling, psychometrics, signal processing, linguistics, and many other areas [17, 27, 5].

1.1. Tensor dimension reduction. Although there are now some nice results for low-rank tensor dimension reduction, the majority of the work (see, e.g., [48, 39, 54, 60]) gives theoretical guarantees for dimensional reducing projections that act on tensors via their matricizations or vectorizations. Two prominent examples are the tensor random projections (TRP) algorithm [53], which is based on the Khatri–Rao product of many smaller random projection maps, and TensorSketch [45, 46], which is based on the tensorization of the CountSketch matrix approach [18]. However, as mentioned above, these methods do not respect the multimodal structure of the tensor (one newer version of TensorSketch that actually does is based on the Tucker format [52]), and the theoretical guarantees are not as general as one would desire: TRP was proved only for tensors of order 2, and the TensorSketch method is mostly applicable to polynomial kernels, that is, a very special case of rank-1 tensors when all the component vectors are copies of the same vector (e.g., [46, 6, 3]).

There are many motivating application areas that utilize efficient tensor dimension reduction, including the acceleration and improvement of machine learning algorithms [46, 39, 50, 57] and finding tensor decompositions (an extensive review of the tensor dimension reduction techniques for low-rank tensor decompositions is given in [42]). Other practical application areas include video and longitudinal imaging [40, 11] and differential equations [10, 41].

Here, our goal is to provide theoretical guarantees for projections that act directly on the tensors themselves without the need for unfolding or vectorization. In particular, this means the projections can be defined *modewise* using the CP-decomposition and that the low-dimensional representations are also tensors but are not vectors. This extends the application for such embeddings to those that cannot afford to perform unfoldings or for which it is not natural to do so. In particular, for tensors in $\mathbb{C}^{n^d}$ for large n and d, this avoids having to store an often impossibly large $m \times n^d$ linear map. In the next section, we elaborate on our main contributions.

We also would like to acknowledge several papers that appeared during the latest stages of preparation of this work and its initial review process. These include the theoretical guarantees for the TRP method for low-rank CP and tensor-train (TT) tensors [47], the new and considerably more efficient algorithm for computing linear
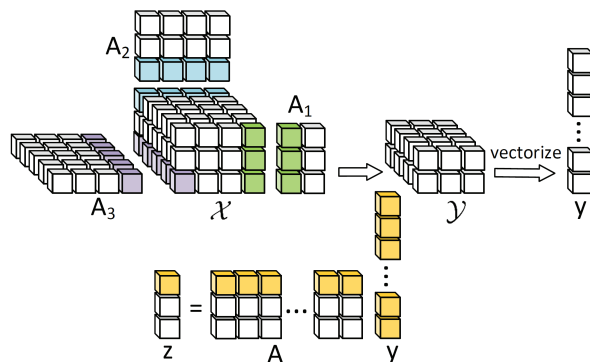
FIG. 1. *An example of 2-stage JL embedding applied to a 3-dimensional tensor $\mathcal{X} \in \mathbb{R}^{3 \times 4 \times 5}$. The output of the first stage is the projected tensor $\mathcal{Y} = \mathcal{X} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3$, where $\mathbf{A}_j$ are JL matrices for $j \in \{1, 2, 3\}$, $\mathbf{A}_1 \in \mathbb{R}^{2 \times 3}$, $\mathbf{A}_2 \in \mathbb{R}^{3 \times 4}$, and $\mathbf{A}_3 \in \mathbb{R}^{4 \times 5}$, resulting in $\mathcal{Y} \in \mathbb{R}^{2 \times 3 \times 4}$. Matching colors have been used to show how the rows of $\mathbf{A}_j$ interact with the mode-j fibers of $\mathcal{X}$ (and the intermediate partially compressed tensors) to generate the elements of the mode-j unfolding of the result after each j-mode product. Next, the resulting tensor is vectorized (leading to $\mathbf{y} \in \mathbb{R}^{24}$), and a second stage JL is then performed to obtain $\mathbf{z} = \mathbf{A}\mathbf{y}$ where $\mathbf{A} \in \mathbb{R}^{3 \times 24}$ and $\mathbf{z} \in \mathbb{R}^3$.*

sketch polynomial kernels [3], and, finally, two works that are most related to our current paper, [32, 43], which show that the Kronecker fast Johnson–Lindenstrauss transform, KFJLT (a special modewise operator based on FFT matrices; see (4.6)) performs a JL-type transform. The first result is more general than the second as it is applied to any tensors, the latter is applicable only to rank-1 tensors, and the efficiency of the compression obtained in these two works is incompatible: the former has better dependence on the dimensions of the original tensor, and the latter has better dependence on the distortion allowed. The second part of our work uses the result of [32] to get an ultimately better result, so further discussion is continued in section 1.2.2. A very nice comparison between these recent results (including our work) is also presented in [43].

1.2. Our contributions. In this paper we analyze modewise tensor embedding strategies for general d-mode tensors. In particular, herein we focus on obliviously embedding an a priori unknown r-dimensional subspace of a given tensor product space $\mathbb{C}^{n_1 \times \cdots \times n_d}$ into a similarly low-dimensional vector space $\mathbb{C}^{\tilde{\mathcal{O}}(r)}$ with high probability. In contrast to the standard approach of effectively vectorizing the tensor product space and then embedding the resulting transformed subspace using standard JL methods involving a single massive $\tilde{\mathcal{O}}(r) \times \prod_{j=1}^{d} n_j$ matrix $\mathbf{M}$ (see, e.g., [39]), the approaches considered herein instead result in the need for generating and storing $d + 1$ significantly smaller matrices $\mathbf{A} \in \mathbb{C}^{\tilde{\mathcal{O}}(r) \times \prod_{\ell=1}^{d} m_\ell}, \mathbf{A}_1 \in \mathbb{C}^{m_1 \times n_1}, \ldots, \mathbf{A}_d \in \mathbb{C}^{m_d \times n_d}$ which are then combined to form a linear embedding operator $L : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{\tilde{\mathcal{O}}(r)}$ via

$$(1.1) \qquad L(\mathcal{X}) := \mathbf{A}\left(\text{vect}\left(\mathcal{X} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\right)\right),$$

where each $\times_j$ is a j-mode product (reviewed below in section 2.1), and vect : $\mathbb{C}^{m_1 \times \cdots \times m_d} \to \mathbb{C}^{\prod_{\ell=1}^{d} m_\ell}$ is a trivial vectorization operator. See Figure 1 for an illustration of how the embedding operator L in (1.1) works in two stages to first map an example 3-mode input tensor $\mathcal{X}$ to a smaller 3-mode tensor $\mathcal{Y}$ and then to a compressed vector $\mathbf{z} = L(\mathcal{X})$.

Let $m' = \tilde{\mathcal{O}}(r)$ be the number of rows one must use for both $\mathbf{M}$ and $\mathbf{A}$ above (as

we shall see, the number of rows required for both matrices will indeed be essentially equivalent). The collective sizes of the matrices needed to define L above will be much smaller (and therefore easier to store, transmit, and generate) than $\mathbf{M}$ whenever $\prod_{\ell=1}^{d} m_\ell + \sum_{\ell=1}^{d} n_\ell \left(\frac{m_\ell}{m'}\right) \ll \prod_{j=1}^{d} n_j$ holds. As a result, much of our discussion below will revolve around bounding the dominant $\prod_{\ell=1}^{d} m_\ell$ term on the left-hand side above, which will also occasionally be referred to as the *intermediate embedding dimension* below. We are now prepared to discuss our two main results.

1.2.1. General oblivious subspace embedding results for low-rank tensor subspaces satisfying an incoherence condition. The first of our results provides new oblivious subspace embeddings for tensor subspaces spanned by bases of rank-1 tensors and establishes related least squares embedding results of value in, e.g., the fitting of a general tensor with an accurate low-rank CP decomposition (CPD) approximation. One of its main contributions is the generality with which it allows one to select the matrices $\mathbf{A}, \mathbf{A}_1, \ldots, \mathbf{A}_d$ used to construct the JL embedding L in (1.1). In particular, it allows each of these matrices to be drawn independently from any nearly optimal family of JL embeddings (as defined immediately below) that the user prefers.

DEFINITION 1 (ε-JL embedding). *Let $\varepsilon \in (0,1)$. We will call a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ an ε-JL embedding of a set $S \subset \mathbb{C}^n$ into $\mathbb{C}^m$ if*

$$\|\mathbf{A}\mathbf{x}\|_2^2 = (1 + \varepsilon_{\mathbf{x}})\|\mathbf{x}\|_2^2$$

holds for some $\varepsilon_{\mathbf{x}} \in (-\varepsilon, \varepsilon)$ for all $\mathbf{x} \in S$.

DEFINITION 2. *Fix $\eta \in (0, 1/2)$, and let $\left\{\mathcal{D}_{(m,n)}\right\}_{(m,n)\in\mathbb{N}\times\mathbb{N}}$ be a family of probability distributions where each $\mathcal{D}_{(m,n)}$ is a distribution over $m \times n$ matrices. We will refer to any such family of distributions as being an η-optimal family of JL embedding distributions if there exists an absolute constant $C \in \mathbb{R}^+$ such that, for any given $\varepsilon \in (0,1)$, $m, n \in \mathbb{N}$ with $m < n$, and nonempty set $\mathcal{S} \subset \mathbb{C}^n$ of cardinality*

$$|S| \leq \eta \exp\left(\frac{\varepsilon^2 m}{C}\right),$$

a matrix $\mathbf{A} \sim \mathcal{D}_{(m,n)}$ will be an ε-JL embedding of $\mathcal{S}$ into $\mathbb{C}^m$ with probability at least $1 - \eta$.

In fact many η-optimal families of JL embedding distributions exist for any given $\eta \in (0, 1/2)$, including, e.g., those associated with random matrices having independent and identically distributed (i.i.d.) sub-Gaussian entries (see Lemma 9.35 in [23]) as well as those associated with sparse Johnson–Lindenstrauss transform (JLT) constructions [34]. The next theorem proves that any desired combination of such matrices can be used to construct a JL embedding L as per (1.1) for any tensor subspace spanned by a basis of rank-1 tensors satisfying an easily testable (and relatively mild[1]) coherence condition. We utilize the notation set forth below in section 2.

THEOREM 1. *Fix $\varepsilon, \eta \in (0, 1/2)$ and $d \geq 3$. Let $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, $n := \max_j n_j \geq 4r + 1$, and let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ spanned by a basis of*

---

[1]In fact the coherence condition required by Theorem 1 will be satisfied by a generic basis of rank-1 tensors with high probability (see section 3.2). Coherence results similar to those presented in section 3.2 have also recently been considered for random tensors in more general parameter regimes by Vershynin [59].

rank-1 *tensors* $\mathcal{B} := \{\bigcirc_{\ell=1}^{d} \mathbf{y}_k^{(\ell)} \mid k \in [r]\}$ *(where* $\bigcirc$ *denotes a tensor outer product operator; see* (2.3) *below) with modewise coherence satisfying*

$$\mu_{\mathcal{B}}^{d-1} := \left(\max_{\ell \in [d]} \max_{k,h \in [r], k \neq h} \left| \left\langle \mathbf{y}_k^{(\ell)}, \mathbf{y}_h^{(\ell)} \right\rangle \right| \right)^{d-1} < 1/2r.$$

Then, one can construct a linear operator $L : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m'}$ *as per* (1.1) *with* $m' \leq C'r \cdot \varepsilon^{-2} \cdot \ln(\frac{47}{\varepsilon \sqrt[d]{\eta}})$ *for an absolute constant* $C' \in \mathbb{R}^+$ *so that with probability at least* $1 - \eta$,

$$\left| \|L(\mathcal{X} - \mathcal{Y})\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right| \leq \varepsilon \|\mathcal{X} - \mathcal{Y}\|^2 \tag{1.2}$$

will hold for all $\mathcal{Y} \in \mathcal{L}$.

If $\mathcal{X} \notin \mathcal{L}$, *the intermediate embedding dimension can be bounded above by*

$$\prod_{\ell=1}^{d} m_\ell \ \leq \ C^d \cdot r^d d^{3d} / \varepsilon^{2d} \cdot \ln^d (n / \sqrt[d]{\eta}) \tag{1.3}$$

for an absolute constant $C \in \mathbb{R}^+$. *If, however,* $\mathcal{X} \in \mathcal{L}$, *then* (1.2) *holds for all* $r < 1/2\mu_{\mathcal{B}}^{d-1}$, *and*

$$\prod_{\ell=1}^{d} m_\ell \ \leq \ \tilde{C}^d \cdot r^2 (d/\varepsilon)^{2d} \cdot \ln^d (2r^2 d/\eta) \tag{1.4}$$

can be achieved, where $\tilde{C} \in \mathbb{R}^+$ *is another absolute constant.*

Proof sketch for Theorem 1. This is largely a restatement of Theorem 4. When defining $L : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m'}$ as per (1.1) following Theorem 4 one should draw $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ with $m_j \geq C_j \cdot rd^3/\varepsilon^2 \cdot \ln(n/\sqrt[d]{\eta})$ from an $(\eta/4d)$-optimal family of JL embedding distributions for each $j \in [d]$, where each $C_j \in \mathbb{R}^+$ is an absolute constant. Furthermore, $\mathbf{A} \in \mathbb{C}^{m' \times \prod_{\ell=1}^{d} m_\ell}$ should be drawn from an $(\eta/2)$-optimal family of JL embedding distributions with m' as above. The probability bound, together with (1.3), then follows. The achievable intermediate embedding dimension when $\mathcal{X} \in \mathcal{L}$ in (1.4) can be obtained from Theorem 3 since the bound $\prod_{\ell=1}^{d} m_\ell \ \leq \ \prod_{\ell=1}^{d} \tilde{C}_\ell \cdot r^{2/d} d^2/\varepsilon^2 \cdot \ln(2r^2 d/\eta)$ can then be utilized in that case. □

One can vectorize the tensors and tensor spaces considered in Theorem 1 using variants of (2.7) to achieve subspace embedding results for subspaces spanned by basis vectors with special Kronecker structure as considered in, e.g., two other recent papers that appeared during the preparation of this article [32, 43]. The most recent of these papers also produces bounds on what amounts to the intermediate embedding dimension of a JL subspace embedding along the lines of (1.1) when $\mathcal{X} \in \mathcal{L}$ (see Theorem 4.1 in [43]). Comparing (1.4) to that result we can see that Theorem 1 has reduced the r dependence of the effective intermediate embedding dimension achieved therein from r^{d+1} to r^2 (now independent of d) for a much more general set of modewise embeddings. However, Theorem 1 incurs a worse dependence on epsilon and needs the stated coherence assumption concerning $\mu_{\mathcal{B}}$ to hold. As a result, Theorem 1 provides a large new class of modewise subspace embeddings that will also have fewer rows than in [43] for a large range of ranks r provided that $\mu_{\mathcal{B}}$ is sufficiently small and ε is sufficiently large.

Note further that the form of (1.2) also makes Theorem 1 useful for solving least squares problems of the type encountered while computing approximate CP decompositions for an arbitrary tensor $\mathcal{X} \notin \mathcal{L}$ using alternating least squares methods (see, e.g., section 4 for a related discussion as well as [9] where modewise strategies were shown to work well for solving such problems in practice). Comparing Theorem 1 to the recent least squares result of the same kind proven in [32] (see Corollary 2.4 there) we can see that Theorem 1 has reduced the r dependence of the effective intermediate embedding dimension achievable in [32] from r^{2d} therein to r^d in (1.3) for a much more general set of modewise embeddings. In exchange, Theorem 1 again incurs a worse dependence on epsilon and needs the stated coherence assumption concerning $\mu_\mathcal{B}$ to hold, however. As a result, Theorem 1 guarantees that a larger class of modewise JL embeddings can be used in least squares applications and that they will also have smaller intermediate embedding dimensions as long as $\mu_\mathcal{B}$ is sufficiently small and ε is sufficiently large.

1.2.2. Fast oblivious subspace embedding results for arbitrary tensor subspaces. Our second main result builds on Theorem 2.1 of Jin, Kolda, and Ward in [32] to provide improved fast subspace embedding results for arbitrary tensor subspaces (i.e., for low-dimensional tensor subspaces whose basis tensors have arbitrary rank and coherence). Let $N := \prod_{j=1}^{d} n_j$. By combining elements of the proof of Theorem 1 with the optimal ε-dependence of Theorem 2.1 in [32], we are able to provide a fast modewise oblivious subspace embedding L as per (1.1) that will simultaneously satisfy (1.2) for all $\mathcal{Y}$ in an entirely arbitrary r-dimensional tensor subspace $\mathcal{L}$ with probability at least $1 - \eta$ while also achieving an intermediate embedding dimension bounded above by

$$(1.5) \qquad C^d \left(\frac{r}{\varepsilon}\right)^2 \cdot \log^{2d-1}\left(\frac{N}{\eta}\right) \cdot \log^4\left(\frac{\log\left(\frac{N}{\eta}\right)}{\varepsilon}\right) \cdot \log N.$$

Above, $C > 0$ is an absolute constant. Note that neither r nor ε in (1.5) is raised to a power of d, which marks a tremendous improvement over all of the previously discussed results when d is large. See Theorem 7 for details.

As alluded to above, the results herein can also be used to create new JL subspace embeddings in the traditional vector space setting. Our next and final main result does this explicitly for arbitrary vector subspaces by restating a variant of Theorem 7 in that context. We expect that this result may be of independent interest outside of the tensor setting.

THEOREM 2. *Fix $\varepsilon, \eta \in (0, 1/2)$ and $d \geq 2$. Let $\mathbf{x} \in \mathbb{C}^N$ such that $\sqrt[d]{N} \in \mathbb{N}$ and $N \geq 4C'/\eta > 1$ for an absolute constant $C' > 0$, and let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^N$ for $\max\left(2r^2 - r, 4r\right) \leq N$. Then, one can construct a random matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with*

$$(1.6) \qquad m \leq C\left[r \cdot \varepsilon^{-2} \cdot \log\left(\frac{47}{\varepsilon \sqrt[r]{\eta}}\right) \cdot \log^4\left(\frac{r \log\left(\frac{47}{\varepsilon \sqrt[r]{\eta}}\right)}{\varepsilon}\right) \cdot \log N\right]$$

for an absolute constant $C > 0$ such that with probability at least $1 - \eta$ it will be the case that

$$\left| \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2^2 - \|\mathbf{x} - \mathbf{y}\|_2^2 \right| \leq \varepsilon \|\mathbf{x} - \mathbf{y}\|_2^2$$

holds for all $\mathbf{y} \in \mathcal{L}$. *Furthermore, the matrix* $\mathbf{A}$ *requires only*

$$(1.7) \qquad \mathcal{O}\left(C_1^d \left(\frac{r}{\varepsilon}\right)^2 \cdot \log^{2d-1}\left(\frac{N}{\eta}\right) \cdot \log^4\left(\frac{\log\left(\frac{N}{\eta}\right)}{\varepsilon}\right) \cdot \log^2 N + d\sqrt[d]{N} \right)$$

random bits and memory for storage for an absolute constant $C_1 > 0$ *and can be multiplied against any vector in just* $\mathcal{O}\left(N \log N\right)$*-time.*

Note that choosing $\mathbf{x} = \mathbf{0}$ produces an oblivious subspace embedding result for $\mathcal{L}$ and that choosing $\mathcal{L}$ to be the column space of a rank-r matrix produces a result useful for least squares sketching.

Proof sketch for Theorem 2. This follows from Theorem 7 after identifying $\mathbb{C}^N$ with $\mathbb{C}^{\sqrt[d]{N} \times \cdots \times \sqrt[d]{N}}$, i.e., after effectively reshaping any given vectors $\mathbf{x}, \mathbf{y}$ under consideration into d-mode tensors $\mathcal{X}, \mathcal{Y}$. Note further that if $\sqrt[d]{N} \notin \mathbb{N}$, then one can implicitly pad the vectors of interest with zeros until $\sqrt{N} \in \mathbb{N}$ (i.e., effectively trivially embedding $\mathbb{C}^N$ into $\mathbb{C}^{\lceil \sqrt[d]{N} \rceil^d}$) before proceeding. □

1.3. Organization. The remainder of the paper is organized as follows. Section 2 provides background and notation for tensors (section 2 and subsection 2.1) and for JL embeddings (subsection 2.2).

We start section 3 with the definitions of the rank of the tensor (and low-rank tensor subspaces) and the maximal modewise coherence of tensor subspace bases. Then we work our way to Theorem 3 which constructs oblivious tensor subspace embeddings via modewise tensor products (for any fixed subspace having low enough modewise coherence). This result is very general in terms of JL embedding maps that one can use as building blocks in each mode. Finally, in subsection 3.2 we discuss the assumption of modewise incoherence and provide several natural examples of incoherent tensor subspaces.

In section 4, we describe the fitting problem for approximately low-rank tensors and explain how modewise dimension reduction (as presented in section 3) reduces the complexity of the problem. Then we build the machinery to show that the solution of the reduced problem will be a good solution for the original problem (in Theorem 4). We conclude section 4 by introducing a two-step embedding procedure that allows one to further reduce the final embedding dimension (this is our second main embedding result, Theorem 7). This improved procedure relies on a specific form of JL embedding of each mode. Both embedding results can be applied to the fitting problem.

In section 5 we present some simple experiments confirming our theoretical guarantees, and then we conclude in section 6.

2. Notation, tensor basics, and linear Johnson–Lindenstrauss embeddings. Tensors, matrices, vectors, and scalars are denoted in different typefaces for clarity below. Calligraphic boldface capital letters are always used for tensors, roman boldface capital letters stand for matrices, roman boldface lowercase letters are used for vectors, and regular roman (lowercase or capital) letters represent scalars. The matrix $\mathbf{I}$ will always represent the identity matrix. The set of the first d natural numbers will be denoted by $[d] := \{1, \ldots, d\}$ for all $d \in \mathbb{N}$.

Throughout the paper, $\otimes$ denotes the Kronecker product of vectors or matrices, and $\bigcirc$ denotes the tensor outer product of vectors or tensors.[2] The symbol $\circ$, on the other hand, represents the composition of functions (see, e.g., section 4). Numbers

---

[2]As (2.3) suggests, it can be applied to tensors with an arbitrary number of modes.

in parentheses used as a subscript or superscript on a tensor denote either *unfoldings* (introduced in section 2.1) when appearing in a subscript or an element in a sequence when appearing in a superscript. The notation $\otimes_{\ell \neq j} \mathbf{v}^{(\ell)}$ for a given set of vectors $\left\{\mathbf{v}^{(\ell)}\right\}_{\ell=1}^{d}$ will always denote the vector $\mathbf{v}^{(d)} \otimes \cdots \otimes \mathbf{v}^{(j+1)} \otimes \mathbf{v}^{(j-1)} \otimes \cdots \otimes \mathbf{v}^{(1)}$. Additional tensor definitions and operations are reviewed below (see, e.g., [36, 21, 56, 61] for additional details and discussion).

2.1. Tensor basics. The set of all d-mode tensors $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ forms a vector space over the complex numbers when equipped with componentwise addition and scalar multiplication. The inner product of $\mathcal{X}, \mathcal{Y} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ will be given by

$$(2.1) \qquad \langle \mathcal{X}, \mathcal{Y} \rangle := \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \mathcal{X}_{i_1,i_2,\ldots,i_d} \overline{\mathcal{Y}_{i_1,i_2,\ldots,i_d}}.$$

This inner product then gives rise to the standard Euclidean norm

$$(2.2) \qquad \|\mathcal{X}\| := \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle} = \sqrt{\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} |\mathcal{X}_{i_1,i_2,\ldots,i_d}|^2}.$$

If $\langle \mathcal{X}, \mathcal{Y} \rangle = 0$, we say that $\mathcal{X}$ and $\mathcal{Y}$ are *orthogonal*. If $\mathcal{X}$ and $\mathcal{Y}$ are orthogonal and also have unit norm (i.e., have $\|\mathcal{X}\| = \|\mathcal{Y}\| = 1$), we say that they are *orthonormal*.

Tensor outer products. The *tensor outer product* of two tensors $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ and $\mathcal{Y} \in \mathbb{C}^{n_1' \times n_2' \times \cdots \times n_{d'}'}$, $\mathcal{X} \bigcirc \mathcal{Y} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d \times n_1' \times n_2' \times \cdots \times n_{d'}'}$ is a $(d+d')$-mode tensor whose entries are given by

$$(2.3) \qquad (\mathcal{X} \bigcirc \mathcal{Y})_{i_1,\ldots,i_d,i_1',\ldots,i_{d'}'} = \mathcal{X}_{i_1,\ldots,i_d} \mathcal{Y}_{i_1',\ldots,i_{d'}'}.$$

Note that when $\mathcal{X}$ and $\mathcal{Y}$ are both vectors, the tensor outer product will reduce to the standard outer product. Some additional standard properties are also listed below in Lemma 1.

Fibers. Let tensor $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_{j-1} \times n_j \times n_{j+1} \times \cdots \times n_d}$. The vectors in $\mathbb{C}^{n_j}$ obtained by fixing all of the indices of $\mathcal{X}$ except for the one that corresponds to its jth mode are called its *mode-j fibers*. Note that any such $\mathcal{X}$ will have $\prod_{\ell \neq j} n_\ell$ mode-j fibers denoted by $\mathcal{X}_{i_1,\ldots,i_{j-1},:,i_{j+1},\ldots,i_d} \in \mathbb{C}^{n_j}$.

Tensor matricization (unfolding). The process of reordering the elements of the tensor into a matrix is known as matricization or unfolding. The mode-j matricization of a tensor $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ is denoted as $\mathbf{X}_{(j)} \in \mathbb{C}^{n_j \times \prod_{m \neq j} n_m}$ and is obtained by arranging $\mathcal{X}$'s mode-j fibers to be the columns of the resulting matrix.

j-mode products: The *j-mode product* of a d-mode tensor

$$\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_{j-1} \times n_j \times n_{j+1} \times \cdots \times n_d}$$

with a matrix $\mathbf{U} \in \mathbb{C}^{m_j \times n_j}$ is another d-mode tensor $\mathcal{X} \times_j \mathbf{U} \in \mathbb{C}^{n_1 \times \cdots \times n_{j-1} \times m_j \times n_{j+1} \times \cdots \times n_d}$. Its entries are given by

$$(2.4) \qquad (\mathcal{X} \times_j \mathbf{U})_{i_1,\ldots,i_{j-1},\ell,i_{j+1},\ldots,i_d} = \sum_{i_j=1}^{n_j} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} \mathbf{U}_{\ell,i_j}$$

for all $(i_1,\ldots,i_{j-1},\ell,i_{j+1},\ldots,i_d) \in [n_1] \times \cdots \times [n_{j-1}] \times [m_j] \times [n_{j+1}] \times \cdots \times [n_d]$. Looking at the mode-$j$ unfoldings of $\mathcal{X} \times_j \mathbf{U}$ and $\mathcal{X}$, one can easily see that their

mode-j matricization can be computed as a regular matrix product

$$(2.5) \qquad (\mathcal{X} \times_j \mathbf{U})_{(j)} = \mathbf{U}\mathbf{X}_{(j)}$$

for all $j \in [d]$. The following simple lemma formally lists several important properties of tensor outer products and modewise products. The proof of Lemma 1 can be found in Appendix A.

LEMMA 1. *Let* $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, $\mathcal{C}, \mathcal{D} \in \mathbb{C}^{n_1' \times n_2' \times \cdots \times n_{d'}'}$, $\alpha, \beta \in \mathbb{C}$, *and* $\mathbf{U}_\ell, \mathbf{V}_\ell \in \mathbb{C}^{m_\ell \times n_\ell}$ *for all* $\ell \in [d]$. *The following properties hold:*
  (i) $(\alpha\mathcal{A} + \beta\mathcal{B}) \bigcirc \mathcal{C} = \alpha\mathcal{A} \bigcirc \mathcal{C} + \beta\mathcal{B} \bigcirc \mathcal{C} = \mathcal{A} \bigcirc \alpha\mathcal{C} + \mathcal{B} \bigcirc \beta\mathcal{C}.$
  (ii) $\langle \mathcal{A} \bigcirc \mathcal{C}, \mathcal{B} \bigcirc \mathcal{D} \rangle = \langle \mathcal{A}, \mathcal{B} \rangle \langle \mathcal{C}, \mathcal{D} \rangle.$
  (iii) $(\alpha\mathcal{A} + \beta\mathcal{B}) \times_j \mathbf{U}_j = \alpha (\mathcal{A} \times_j \mathbf{U}_j) + \beta (\mathcal{B} \times_j \mathbf{U}_j).$
  (iv) $\mathcal{A} \times_j (\alpha\mathbf{U}_j + \beta\mathbf{V}_j) = \alpha (\mathcal{A} \times_j \mathbf{U}_j) + \beta (\mathcal{A} \times_j \mathbf{V}_j).$
  (v) *If* $j \neq \ell$, *then* $\mathcal{A} \times_j \mathbf{U}_j \times_\ell \mathbf{V}_\ell = (\mathcal{A} \times_j \mathbf{U}_j) \times_\ell \mathbf{V}_\ell = (\mathcal{A} \times_\ell \mathbf{V}_\ell) \times_j \mathbf{U}_j = \mathcal{A} \times_\ell \mathbf{V}_\ell \times_j \mathbf{U}_j.$
  (vi) *If* $W \in \mathbb{C}^{p \times m_j}$, *then* $\mathcal{A} \times_j \mathbf{U}_j \times_j \mathbf{W} = (\mathcal{A} \times_j \mathbf{U}_j) \times_j \mathbf{W} = \mathcal{A} \times_j (\mathbf{W}\mathbf{U}_j) = \mathcal{A} \times_j \mathbf{W}\mathbf{U}_j.$

A generalization of the observation (2.5) is available: unfolding the tensor

$$(2.6) \qquad \mathcal{Y} = \mathcal{X} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_d \mathbf{U}^{(d)} =: \mathcal{X} \underset{j=1}{\overset{d}{\times}} \mathbf{U}^{(j)}$$

along the jth mode is equivalent to

$$(2.7) \qquad \mathbf{Y}_{(j)} = \mathbf{U}^{(j)}\mathbf{X}_{(j)} \left(\mathbf{U}^{(d)} \otimes \cdots \mathbf{U}^{(j+1)} \otimes \mathbf{U}^{(j-1)} \cdots \otimes \mathbf{U}^{(1)} \right)^\top,$$

where $\otimes$ is the matrix Kronecker product (see [36]). In particular, (2.7) implies that the matricization $\left(\mathcal{X} \times_j \mathbf{U}^{(j)}\right)_{(j)} = \mathbf{U}^{(j)}\mathbf{X}_{(j)}$.[3] On a related note, one can also express the relation between the vectorized forms of $\mathcal{X}$ and $\mathcal{Y}$ in (2.6) as

$$(2.8) \qquad \text{vect}(\mathcal{Y}) = \left(\mathbf{U}^{(d)} \otimes \cdots \otimes \mathbf{U}^{(1)} \right) \text{vect}(\mathcal{X}),$$

where $\text{vect}(\cdot)$ is the vectorization operator.

Finally, it is also worth noting that trivial inner product preserving isomorphisms exist between a tensor space $\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ and any of its matricized versions (i.e., mode-j matricization can be viewed as an isomorphism between the original tensor vector space $\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ and its mode-j matricized target vector space $\mathbb{C}^{n_j \times \prod_{m \neq j} n_m}$). In particular, the process of matricizing tensors is linear. If, for example, $\mathcal{X}, \mathcal{Y} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, then one can see that the mode-j matricization of $\mathcal{X} + \mathcal{Y} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ is $(\mathcal{X} + \mathcal{Y})_{(j)} = \mathbf{X}_{(j)} + \mathbf{Y}_{(j)}$ for all modes $j \in [d]$.

2.2. Linear Johnson–Lindenstrauss embeddings. Many linear ε-JL embedding matrices exist [33, 2, 20, 37, 38], with the best achievable target dimension being $m = \mathcal{O}(\log(|S|)/\varepsilon^2)$ for arbitrary S (see [38] for results concerning the optimality of this embedding dimension). Of course, one can define JL embedding on tensors in a similar way, namely, as linear maps approximately preserving the tensor norm.

---

[3]Simply set $\mathbf{U}^{(m)} = \mathbf{I}$ (the identity) for all $m \neq n$ in (2.7). This fact also easily follows directly from the definition of the j-mode product.

DEFINITION 3 (tensor ε-JL embedding). *A linear operator* $L : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to$ $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$ *is an ε-JL embedding of a set* $S \subset \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ *into* $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$ *if*

$$\|L(\mathcal{X})\|^2 = (1 + \varepsilon_{\mathcal{X}}) \|\mathcal{X}\|^2$$

holds for some $\varepsilon_{\mathcal{X}} \in (-\varepsilon, \varepsilon)$ *for all* $\mathcal{X} \in S$.

It is easy to check that JL embeddings can preserve pairwise inner products. The proof of the following lemma can be found in Appendix A.

LEMMA 2. *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, *and suppose that* $\mathbf{A} \in \mathbb{C}^{m \times n}$ *is an ε-JL embedding of the vectors*

$$\{\mathbf{x} - \mathbf{y}, \mathbf{x} + \mathbf{y}, \mathbf{x} - \mathrm{i}\mathbf{y}, \mathbf{x} + \mathrm{i}\mathbf{y}\} \subset \mathbb{C}^n$$

into $\mathbb{C}^m$. *Then,*

$$|\langle \mathbf{Ax},\ \mathbf{Ay} \rangle - \langle \mathbf{x},\ \mathbf{y} \rangle| \ \leq\ 2\varepsilon \left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 \right) \ \leq\ 4\varepsilon \cdot \max \left\{ \|\mathbf{x}\|_2^2, \|\mathbf{y}\|_2^2 \right\}.$$

Moreover, if $\mathcal{X}, \mathcal{Y} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, *and assuming that L is an ε-JL embedding of the tensors*

$$\{\mathcal{X} - \mathcal{Y}, \mathcal{X} + \mathcal{Y}, \mathcal{X} - \mathrm{i}\mathcal{Y}, \mathcal{X} + \mathrm{i}\mathcal{Y}\} \subset \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$$

into $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$, *then,*

$$|\langle L(\mathcal{X}),\ L(\mathcal{Y}) \rangle - \langle \mathcal{X},\ \mathcal{Y} \rangle| \ \leq\ 2\varepsilon \left(\|\mathcal{X}\|^2 + \|\mathcal{Y}\|^2 \right) \ \leq\ 4\varepsilon \cdot \max \left\{ \|\mathcal{X}\|^2, \|\mathcal{Y}\|^2 \right\}.$$

In the case where a more general set S is embedded using JL embeddings, for example, a low-rank subspace of tensors, in order to pass to a smaller finite set, a discretization technique can be used. Due to linearity, it actually suffices to discretize the unit ball of the space in question. In the next lemma we present a simple subspace embedding result based on a standard covering argument (see, e.g., [7, 23]). We include its proof in Appendix A for the sake of completeness.

LEMMA 3. *Fix* $\varepsilon \in (0, 1)$. *Let* $\mathcal{L}$ *be an r-dimensional subspace of* $\mathbb{C}^n$, *and let* $\mathcal{C} \subset \mathcal{L}$ *be an $(\varepsilon/16)$-net of the $(r-1)$-dimensional Euclidean unit sphere* $\mathcal{S}_{\ell^2} \subset \mathcal{L}$. *Then, if* $\mathbf{A} \in \mathbb{C}^{m \times n}$ *is an $(\varepsilon/2)$-JL embedding of* $\mathcal{C}$, *it will also satisfy*

(2.9) $$(1 - \varepsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \varepsilon)\|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \mathcal{L}$. *Furthermore, we note that there exists an $(\varepsilon/16)$-net such that* $|\mathcal{C}| \leq \left(\frac{47}{\varepsilon} \right)^r$.

Remark 1. We will see later in the text that the cardinality $(47/\varepsilon)^r$ (exponential in r) can be too big to produce tensor JL embeddings with optimal embedding dimensions. In this case one can use a much coarser "discretization" to improve the dependence on r based on, e.g., the next lemma. We point out that there is indeed a trade-off when using an approach such as Lemma 4; instead of controlling the norms of all vectors in a subspace by embedding a cover of the unit ball as in Lemma 3, in Lemma 4 we instead control the norms of all vectors in a subspace by embedding an orthonormal basis that approximately preserves their orthogonality. The trade-off is that one needs to preserve the angles between the orthonormal basis vectors quite accurately in order to ensure that all of the vectors in their span also have their norms preserved well as a result.

With Lemma 2 in hand we are now able to prove a secondary subspace embedding result, which, though it leads to suboptimal results in the vector setting, will be valuable for higher mode tensors.

LEMMA 4. *Fix $\varepsilon \in (0,1)$, and let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ spanned by a set of r orthonormal basis tensors $\{\mathcal{T}_k\}_{k \in [r]}$. If L is an $(\varepsilon/4r)$-JL embedding of the $4\binom{r}{2} + r = 2r^2 - r$ tensors*

$$\left(\bigcup_{1 \leq h < k \leq r} \{\mathcal{T}_k - \mathcal{T}_h, \mathcal{T}_k + \mathcal{T}_h, \mathcal{T}_k - \mathrm{i}\mathcal{T}_h, \mathcal{T}_k + \mathrm{i}\mathcal{T}_h\} \right) \bigcup \{\mathcal{T}_k\}_{k \in [r]} \subset \mathcal{L}$$

into $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$, then

$$\left| \|L(\mathcal{X})\|^2 - \|\mathcal{X}\|^2 \right| \leq \varepsilon \|\mathcal{X}\|^2$$

holds for all $\mathcal{X} \in \mathcal{L}$.

Proof. Appealing to Lemma 2 we can see that $|\varepsilon_{k,h}| := |\langle L(\mathcal{T}_k), L(\mathcal{T}_h) \rangle - \langle \mathcal{T}_k, \mathcal{T}_h \rangle| \leq \varepsilon/r$ for all $h, k \in [r]$. As a consequence, we have for any $\mathcal{X} = \sum_{k=1}^r \alpha_k \mathcal{T}_k \in \mathcal{L}$ that

$$\left| \|L(\mathcal{X})\|^2 - \|\mathcal{X}\|^2 \right| = \left| \sum_{k=1}^r \sum_{h=1}^r \alpha_k \overline{\alpha_h} \left(\langle L(\mathcal{T}_k), L(\mathcal{T}_h) \rangle - \langle \mathcal{T}_k, \mathcal{T}_h \rangle \right) \right| = \left| \sum_{k=1}^r \sum_{h=1}^r \alpha_k \overline{\alpha_h} \varepsilon_{k,h} \right|$$

$$\leq \sum_{k=1}^r |\alpha_k| \sum_{h=1}^r |\alpha_h| \, |\varepsilon_{k,h}| \leq \sum_{k=1}^r |\alpha_k| \, \|\boldsymbol{\alpha}\|_2 \left(\frac{\varepsilon}{\sqrt{r}} \right) \leq \varepsilon \|\boldsymbol{\alpha}\|_2^2.$$

To finish we now note that $\|\mathcal{X}\|^2 = \|\boldsymbol{\alpha}\|_2^2$ due to the orthonormality of the basis tensors $\{\mathcal{T}_k\}_{k \in [r]}$. ∎

3. Modewise linear Johnson–Lindenstrauss embeddings of low-rank tensors. In this section, we consider low-rank tensor subspace embeddings for tensors with low-rank expansions in terms of rank-1 tensors (i.e., for tensors with low-rank CP decompositions). Our general approach will be to utilize subspace embeddings along the lines of Lemmas 3 and 4 in this setting. However, the fact that our basis tensors are rank-1 will cause us some difficulties. Principally, among those difficulties will be our inability to guarantee that we can find an orthonormal, or even fairly incoherent, basis of rank-1 tensors that span any particular r-dimensional tensor subspace $\mathcal{L}$ we may be interested in below.

Going forward, we will consider the *standard form* of a given rank-r d-mode tensor defined by

$$(3.1) \qquad \mathcal{Y} := \sum_{k=1}^r \alpha_k \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \in \mathbb{C}^{n_1 \times \cdots \times n_d},$$

where the vectors making up the rank-1 basis tensors are normalized so that $\|\mathbf{y}_k^{(\ell)}\|_2 = 1$ for all $\ell \in [d]$ and $k \in [r]$.

Given a set of rank-1 tensors spanning a tensor subspace, one can define the coherence of the basis.

DEFINITION 4 (modewise coherence of a basis of rank-1 tensors). *If a tensor subspace is spanned by a basis of rank-1 tensors $\mathcal{B} := \{\bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \mid k \in [r]\}$ with $\|\mathbf{y}_k^{(\ell)}\|_2 = 1$ for all $\ell \in [d]$ and $k \in [r]$, we denote the maximum modewise coherence of the basis and the basis coherence by*

$$(3.2) \qquad \mu_{\mathcal{B}} := \max_{\ell \in [d]} \mu_{\mathcal{B},\ell} \quad and \quad \mu_{\mathcal{B}}' := \max_{\substack{k,h \in [r] \\ k \neq h}} \prod_{\ell=1}^d \left| \left\langle \mathbf{y}_k^{(\ell)}, \mathbf{y}_h^{(\ell)} \right\rangle \right|,$$

respectively, where

$$\mu_{\mathcal{B},\ell} := \max_{\substack{k,h \in [r] \\ k \neq h}} \left| \left\langle \mathbf{y}_k^{(\ell)}, \mathbf{y}_h^{(\ell)} \right\rangle \right|$$

is the modewise coherence of the basis for $\ell \in [d]$.

Note that $\mu_{\mathcal{B}}, \mu'_{\mathcal{B}} \in [0,1]$ and that $\mu'_{\mathcal{B}} \leq \prod_{\ell=1}^d \mu_{\mathcal{B},\ell} \leq \mu_{\mathcal{B}}^d$ always hold. Given any tensor $\mathcal{Y}$ in the span of a basis $\mathcal{B}$ of rank-1 tensors, we will also refer (with some abuse of notation) to its modewise coherence and maximum modewise coherence as being equal to the modewise coherence and maximum modewise coherence of the given basis $\mathcal{B}$ defined in Definition 4. That is, we will say that

(3.3) $\mu_{\mathcal{Y},\ell} = \mu_{\mathcal{B},\ell}$ for $\ell \in [d]$,   and $\mu_{\mathcal{Y}} = \mu_{\mathcal{B}}$

for all $\mathcal{Y} \in \mathcal{B}$. Similarly, the basis coherence of any such $\mathcal{Y} \in \mathcal{B}$ will be said to equal the basis coherence also defined in Definition 4, i.e., $\mu'_{\mathcal{Y}} = \mu'_{\mathcal{B}}$. *It should be remembered below, however, that the quantities $\mu_{\mathcal{Y},\ell}$, $\mu_{\mathcal{Y}}$, $\mu'_{\mathcal{Y}}$ always depend on the particular basis $\mathcal{B}$ under consideration.*

The main result of this section is the following oblivious subspace embedding theorem for low-rank tensors.

THEOREM 3. *Fix $\delta, \eta \in (0, 1/2)$ and $d \geq 2$. Let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ spanned by a basis of rank-1 tensors $\mathcal{B} := \{ \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \mid k \in [r] \}$ with modewise coherence (as per (3.2)) satisfying $\mu_{\mathcal{B}}^{d-1} < 1/2r$. For each $j \in [d]$ draw $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ with*

(3.4) $$m_j \geq \tilde{C} \cdot r^{2/d} d^2 / \varepsilon^2 \cdot \ln \left(2r^2 d / \eta \right)$$

from an (η/d)-optimal family of JL embedding distributions, where $\tilde{C} \in \mathbb{R}^+$ is an absolute constant. Then with probability at least $1 - \eta$ we have

(3.5) $$\left| \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2 - \|\mathcal{Y}\|^2 \right| \leq \varepsilon \|\mathcal{Y}\|^2$$

for all $\mathcal{Y} \in \mathcal{L}$.

Remark 2. A modewise incoherence assumption is necessary for our proof of Theorem 3. (Indeed, we initially obtain the upper estimate for the distortion in (3.5) in terms of $\|\alpha\|$ instead of $\|\mathcal{Y}\|$. As suggested by Lemma 7 below, in the case when $\mu_{\mathcal{Y}}$ is large these two norms can be incompatible.) However, numerical experiments with the coherent model show compatible results even for very coherent tensors. See, e.g., Figure 2 and the additional relevant discussion in section 5.

The next subsection presents all the components of the proof of Theorem 3, whereas the details of the auxiliary lemmas and propositions are deferred to Appendix B.

3.1. Proof of the oblivious tensor subspace embedding Theorem 3. The first auxiliary lemma deals with how j-mode products can change the standard form and modewise coherence of a given tensor that lies in a tensor subspace spanned by r rank-1 tensors.

LEMMA 5. *Let $j \in [d]$, let $\mathbf{B} \in \mathbb{C}^{m \times n_j}$, and let $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ be a rank-r tensor as per (3.1) such that $\min_{k \in [r]} \|\mathbf{B}\mathbf{y}_k^{(j)}\|_2 > 0$. Then $\mathcal{Y}' := \mathcal{Y} \times_j \mathbf{B}$ can be written in*

standard form as

$$\mathcal{Y}' = \sum_{k=1}^{r} \alpha_k \left\| \mathbf{B} \mathbf{y}_k^{(j)} \right\|_2 \left(\left(\bigcirc_{\ell < j} \mathbf{y}_k^{(\ell)} \right) \bigcirc \frac{\mathbf{B} \mathbf{y}_k^{(j)}}{\left\| \mathbf{B} \mathbf{y}_k^{(j)} \right\|_2} \bigcirc \left(\bigcirc_{\ell > j}^{d} \mathbf{y}_k^{(\ell)} \right) \right).$$

Furthermore, the j-mode coherence of $\mathcal{Y}'$ as above will satisfy

$$\mu_{\mathcal{Y}',j} = \max_{\substack{k,h \in [r] \\ k \neq h}} \frac{\left| \left\langle \mathbf{B} \mathbf{y}_k^{(j)}, \mathbf{B} \mathbf{y}_h^{(j)} \right\rangle \right|}{\left\| \mathbf{B} \mathbf{y}_k^{(j)} \right\|_2 \left\| \mathbf{B} \mathbf{y}_h^{(j)} \right\|_2}$$

so that

$$\mu_{\mathcal{Y}'} = \max \left(\mu_{\mathcal{Y}',j} , \max_{\ell \in [d] \setminus \{j\}} \max_{\substack{k,h \in [r] \\ k \neq h}} \left| \left\langle \mathbf{y}_k^{(\ell)}, \mathbf{y}_h^{(\ell)} \right\rangle \right| \right).$$

The proofs of this and all subsequent intermediate results stated in this section can be found in Appendix B. The next lemma gives us a useful expression for the norm of a tensor after a *j*-mode product in terms of vector inner products.

LEMMA 6. *Let $j \in [d]$, let $\mathbf{B} \in \mathbb{C}^{m \times n_j}$, and let $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ be a rank-r tensor in standard form as per* (3.1). *Then,*

$$\left\| \mathcal{Y} \times_j \mathbf{B} \right\|^2 = \sum_{k,h=1}^{r} \sum_{a=1}^{\Pi_{\ell \neq j} n_\ell} \alpha_k \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)_a \overline{\alpha_h \left(\otimes_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right)_a} \left\langle \mathbf{B} \mathbf{y}_k^{(j)}, \mathbf{B} \mathbf{y}_h^{(j)} \right\rangle,$$

where $(\mathbf{u})_a$ denotes the ath coordinate of a vector $\mathbf{u}$.

The following proposition demonstrates that a single modewise JL embedding of any low-rank tensor $\mathcal{Y}$ of the form (3.1) will preserve its norm up to an error depending on the overall ℓ^2-norm of its coefficients $\boldsymbol{\alpha} \in \mathbb{C}^r$. In order to accomplish this, we will connect JL embeddings of combinations of the basis vectors $\mathbf{y}_k^{(\ell)}$ to the embedding properties of any low-rank tensor $\mathcal{Y}$ of the form (3.1). We will employ Lemma 2 for this purpose and consider the following sets $\mathcal{S}_j'$ defined using the basis vectors $\mathbf{y}_k^{(\ell)}$. For each mode $j \in [d]$ of any rank-*r* tensor as per (3.1), we can associate the following set $\mathcal{S}_j'$ of $4\binom{r}{2} + r = 2r^2 - r$ vectors in $\mathbb{C}^{n_j}$ that will be of use later together with Lemma 2:
(3.6)

$$\mathcal{S}_j' := \left(\bigcup_{1 \leq h < k \leq r} \left\{ \mathbf{y}_k^{(j)} - \mathbf{y}_h^{(j)}, \mathbf{y}_k^{(j)} + \mathbf{y}_h^{(j)}, \mathbf{y}_k^{(j)} - \mathbb{i} \mathbf{y}_h^{(j)}, \mathbf{y}_k^{(j)} + \mathbb{i} \mathbf{y}_h^{(j)} \right\} \right) \bigcup \left\{ \mathbf{y}_k^{(j)} \right\}.$$

PROPOSITION 1. *Let $j \in [d]$, and let $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ be a rank-r tensor as per* (3.1). *Suppose that $\mathbf{A} \in \mathbb{C}^{m \times n_j}$ is an $(\varepsilon/4)$-JL embedding of all the vectors from the set $\mathcal{S}_j'$ defined as per* (3.6) *into $\mathbb{C}^m$. Let $\mathcal{Y}' := \mathcal{Y} \times_j \mathbf{A}$, and rewrite it in standard form so that*

$$\mathcal{Y}' = \sum_{k=1}^{r} \alpha_k' \left(\left(\bigcirc_{\ell < j} \mathbf{y}_k^{(\ell)} \right) \bigcirc \frac{\mathbf{A} \mathbf{y}_k^{(j)}}{\left\| \mathbf{A} \mathbf{y}_k^{(j)} \right\|_2} \bigcirc \left(\bigcirc_{\ell > j}^{d} \mathbf{y}_k^{(\ell)} \right) \right).$$

Then all of the following hold:

($\dagger$) $|\alpha'_k - \alpha_k| \leq \varepsilon|\alpha_k|/4$ *for all* $k \in [r]$ *so that* $\|\boldsymbol{\alpha}'\|_\infty \leq (1 + \varepsilon/4)\|\boldsymbol{\alpha}\|_\infty$.

($\dagger\dagger$) $\mu_{\mathcal{Y}',j} \leq \frac{\mu_{\mathcal{Y},j}+\varepsilon}{1-\varepsilon/4}$, *and* $\mu_{\mathcal{Y}',\ell} = \mu_{\mathcal{Y},\ell}$ *for all* $\ell \in [d] \setminus \{j\}$.

($\dagger\dagger\dagger$) $\left|\|\mathcal{Y}'\|^2 - \|\mathcal{Y}\|^2\right| \leq \varepsilon(r+1)\|\boldsymbol{\alpha}\|_2^2$.

Note that part ($\dagger\dagger\dagger$) of Proposition 1 bounds $\left|\|\mathcal{Y}'\|^2 - \|\mathcal{Y}\|^2\right|$ with respect to $\|\boldsymbol{\alpha}\|_2^2$. Traditional JL-type error guarantees typically want to prove error bounds of the form $\left|\|\mathcal{Y}'\|^2 - \|\mathcal{Y}\|^2\right| \leq C_\varepsilon\|\mathcal{Y}\|^2$, however. The next lemma bounds $\|\boldsymbol{\alpha}\|_2^2$ by $\|\mathcal{Y}\|^2$ so that the reader who desires such bounds can obtain them easily for any tensor with sufficiently small modewise coherence.

LEMMA 7. *Let* $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ *be a rank-r tensor as per* (3.1) *with the basis coherence* $\mu'_{\mathcal{Y}} < (r-1)^{-1}$. *Then,*

$$\|\boldsymbol{\alpha}\|_2^2 \leq \left(\frac{1}{1-(r-1)\mu'_{\mathcal{Y}}}\right)\|\mathcal{Y}\|^2 \leq \left(\frac{1}{1-(r-1)\prod_{\ell=1}^d \mu_{\mathcal{Y},\ell}}\right)\|\mathcal{Y}\|^2 \leq \left(\frac{1}{1-(r-1)\mu_{\mathcal{Y}}^d}\right)\|\mathcal{Y}\|^2.$$

We are now prepared to establish Proposition 2, our main component of the proof of Theorem 3 in this section. Recall that combining it with Lemma 7 provides traditional JL embedding error bounds.

PROPOSITION 2. *Let* $\varepsilon \in (0, 3/4]$, *let* $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ *be a rank-r tensor expressed in standard form as per* (3.1)*, and let* $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ *be an* $(\varepsilon/4d)$*-JL embedding of all the vectors from the set* $\mathcal{S}'_j$ *defined as per* (3.6) *into* $\mathbb{C}^{m_j}$ *for each* $j \in [d]$. *Then,*

(3.7)
$$\begin{aligned}
\left|\|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2\right| &\leq \varepsilon\left(\mathbb{e} + \mathbb{e}^2\sqrt{r(r-1)} \cdot \max\left(\varepsilon^{d-1}, \mu_{\mathcal{Y}}^{d-1}\right)\right)\|\boldsymbol{\alpha}\|_2^2 \\
&\leq \varepsilon\mathbb{e}^2(r+1)\|\boldsymbol{\alpha}\|_2^2
\end{aligned}$$

always holds. Here, $\mu_{\mathcal{Y}}$ *is the maximum modewise coherence of the tensor defined by* (3.3)*. Furthermore, if* $\mu_{\mathcal{Y}} = 0$, *then*

$$\left|\|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2\right| \leq \left(\varepsilon + \mathbb{e}\sqrt{r(r-1)}\varepsilon^d\right)\mathbb{e}\|\boldsymbol{\alpha}\|_2^2.$$

In addition, Proposition 2 can be extended to show that modewise compression preserves scalar products between two tensors $\mathcal{X}$ and $\mathcal{Y}$ spanned by the same rank-1 tensors.

PROPOSITION 3. *Suppose that both* $\mathcal{X}, \mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ *can be represented in terms of the same basis* $\{\mathbf{y}_k^{(\ell)}\}$ *for* $k = 1, \ldots, r$ *and* $l = 1, \ldots, d$ *in their standard form* (3.1)*. Let* $\varepsilon \in (0, 3/4]$, *and let* $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ *be a* $(\varepsilon/4d)$*-JL embedding of the set* S'_j *defined as per* (3.6) *for each* $j \in [d]$. *Then,*

$$\left|\langle \mathcal{X} \times_{j=1}^d \mathbf{A}_j, \ \mathcal{Y} \times_{j=1}^d \mathbf{A}_j\rangle - \langle\mathcal{X}, \ \mathcal{Y}\rangle\right| \leq 4\varepsilon' \cdot \frac{\max\left\{\|\mathcal{X}\|^2, \|\mathcal{Y}\|^2\right\}}{1-(r-1)\mu'_{\mathcal{Y}}},$$

where

(3.8)
$$\varepsilon' := \begin{cases} \left(\varepsilon + \mathbb{e}\sqrt{r(r-1)}\varepsilon^d\right)\mathbb{e} & \text{if } \mu_{\mathcal{Y}} = 0, \\ \varepsilon\left(\mathbb{e} + \mathbb{e}^2\sqrt{r(r-1)} \cdot \max\left(\varepsilon^{d-1}, \mu_{\mathcal{Y}}^{d-1}\right)\right) & \text{otherwise.} \end{cases}$$

Proof. Using the polarization identity in combination with Lemma 1 and Proposition 2, we can see that

$$
\left| \langle \mathcal{X} \times_{j=1}^d \mathbf{A}_j, \mathcal{Y} \times_{j=1}^d \mathbf{A}_j \rangle - \langle \mathcal{X}, \mathcal{Y} \rangle \right|
$$

$$
= \left| \frac{1}{4} \sum_{\ell=0}^3 \mathbb{i}^\ell \left(\left\| \mathcal{X} \times_{j=1}^d \mathbf{A}_j + \mathbb{i}^\ell \mathcal{Y} \times_{j=1}^d \mathbf{A}_j \right\|_2^2 - \left\| \mathcal{X} + \mathbb{i}^\ell \mathcal{Y} \right\|_2^2 \right) \right|
$$

$$
\leq \frac{1}{4} \sum_{\ell=0}^3 \varepsilon' \left\| \boldsymbol{\beta} + \mathbb{i}^\ell \boldsymbol{\alpha} \right\|_2^2 \ \leq \ \varepsilon' \left(\left\| \boldsymbol{\beta} \right\|_2 + \left\| \boldsymbol{\alpha} \right\|_2 \right)^2
$$

$$
\leq 2\varepsilon' \left(\left\| \boldsymbol{\beta} \right\|_2^2 + \left\| \boldsymbol{\alpha} \right\|_2^2 \right) \ \leq \ 4\varepsilon' \cdot \max \left\{ \left\| \boldsymbol{\beta} \right\|_2^2, \left\| \boldsymbol{\alpha} \right\|_2^2 \right\},
$$

where the second-to-last inequality follows from Young's inequality for products. An application of Lemma 7 yields the final inequality. $\square$

Propositions 2 and 3 guarantee that modewise JL embeddings approximately preserve the norms and inner products between all tensors in the span of the set

$$
\mathcal{B} := \left\{ \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \ \big| \ k \in [r] \right\} \subset \mathbb{C}^{n_1 \times \cdots \times n_d}.
$$

Let

$$
\mathcal{L} := \mathrm{span} \left(\left\{ \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \ \big| \ k \in [r] \right\} \right).
$$

Employing η-optimal JL embeddings (as per Definition 2), we can now prove the main result of this section, Theorem 3.

Proof of Theorem 3. Let $\mathcal{Y} \in \mathcal{L}$. By Proposition 3, the linear operator L defined as $L(\mathcal{Z}) = \mathcal{Z} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d$ is an ε-JL embedding of $\mathcal{Y}$ if
- $4/\left(1 - (r-1)\mu'_{\mathcal{B}} \right) \leq 8$, and
- each $\mathbf{A}_j$ is a $(\delta/4d)$-JL embedding of the set S'_j of cardinality $|S'_j| \leq 2r^2 - r$, where the dependence $\varepsilon'(\delta)$ is defined by (3.8), and $\varepsilon \geq 8\varepsilon'$.

The first condition is satisfied since the basis incoherence condition implies

$$
\mu'_{\mathcal{B}} \leq \mu_{\mathcal{B}}^d \leq 1/2 \left(r - 1 \right).
$$

Hence, $8(1 - (r-1)\mu'_{\mathcal{B}}) \geq 4$. To check the second condition, note that due to (3.8), it is enough to use an ε that satisfies

$$
\varepsilon \geq 8\delta e + 8\delta e^2 r \max \left(\delta^{d-1}, \mu_{\mathcal{B}}^{d-1} \right),
$$

and having $\delta := \varepsilon/16e \cdot (1/r)^{1/d}$ ensures that it does. Finally, if each matrix $\mathbf{A}_j$ is taken from an (η/d)-optimal family of JL distributions, it will be an $(\delta/4d)$-JL embedding of S'_j into $\mathbb{C}^{m_j}$ with probability $1 - \eta/d$ as long as

$$
|S'_j| = 2r^2 - r \leq \frac{\eta}{d} \exp \left(\frac{\delta^2 m_j}{16d^2 C} \right),
$$

which is satisfied for each m_j defined in (3.4). Taking a union bound over all d modes then concludes the proof of Theorem 3. $\square$

Remark 3 (JL-type embeddings for low-rank matrices). Theorem 3 (as well as the above results, including Proposition 2) can be applied in the special case where $\mathcal{X} = \mathbf{X}$ is a matrix in $\mathbb{C}^{n_1 \times n_2}$. In this case, the CP-rank is the usual matrix rank, and

the CP decomposition becomes the regular SVD decomposition of the matrix, which can be computed efficiently in parallel (see, e.g., [28]). In particular, the basis vectors are orthogonal to each other in this case. The result of Theorem 3 implies that by taking A and B as matrices belonging to the $(\eta/2)$-JL embedding family and of sizes $n_1 \times m_1$ and $n_2 \times m_2$, respectively, such that $m_j \gtrsim r \ln(r/\sqrt{\eta})/\varepsilon^2$ (for $j = 1, 2$), we get the following JL-type result for the Frobenius matrix norm: with probability $1 - \eta$,

$$\|A^T \mathbf{X} B\|_F^2 = (1 + \tilde{\varepsilon})\|\mathbf{X}\|_F^2 \quad \text{for some } |\tilde{\varepsilon}| \leq \varepsilon.$$

3.2. Naturally incoherent tensor bases. Again, we remind the reader that Lemma 7 can be used in combination with the theorems and corollaries above/below in order to provide JL embedding results of the usual type. In order for Lemma 7 to apply, however, we need the coherence $\mu'_{\mathcal{B}}$ of the basis $\mathcal{B}$ to satisfy $\mu'_{\mathcal{B}} < (r-1)^{-1}$. One popular set of bases with this property are those that result from considering tensors whose Tucker decompositions [55, 35, 28] have core tensors with a small number of nonzero entries. More specifically, let $\mathcal{C} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, $\mathbf{U}^{(j)} \in \mathbb{C}^{n_j \times n_j}$ be unitary for all $j \in [d]$, and let $\mathcal{S} \subset [n_1] \times \cdots \times [n_d]$ be a set of r indices in $\mathcal{C}$. Now consider the r-dimensional tensor subspace

$$\mathcal{L}_{\text{Tucker}} := \left\{ \mathcal{X} \mid \mathcal{X} = \mathcal{C} \times_{j=1}^d \mathbf{U}^{(j)} \text{ with } \mathcal{C}_{\mathbf{i}} = 0 \text{ for all } \mathbf{i} \notin \mathcal{S} \right\}.$$

One can see that any tensor $\mathcal{Y} \in \mathcal{L}_{\text{Tucker}}$ can be written in standard form as per (3.1) with, for all $\ell \in [d]$, $\mathbf{y}_k^{(\ell)} = \mathbf{U}_{k'}^{(\ell)}$ for some column $k' \in [n_\ell]$. As a result, $\mu'_{\mathcal{Y}} = \mu'_{\mathcal{B}} = 0$ will hold due to the orthogonality of the columns of each $\mathbf{U}^{(\ell)}$ matrix. We therefore have the following special case of Proposition 2 in this setting.

COROLLARY 1. *Suppose that* $\mathcal{Y} \in \mathcal{L}_{\text{Tucker}} \subset \mathbb{C}^{n_1 \times \cdots \times n_d}$. *Let* $\varepsilon \in (0, 3/4]$, *and let* $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ *be defined as per Proposition* 2 *for each* $j \in [d]$. *Then,*

$$\left| \|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2 \right| \leq \varepsilon' \|\mathcal{Y}\|^2,$$

where

$$\varepsilon' := \begin{cases} \left(\varepsilon + \mathrm{e}\sqrt{r(r-1)}\varepsilon^d \right) \mathrm{e} & \text{if } \mu_{\mathcal{B}} = 0, \\ \varepsilon \left(\mathrm{e} + \mathrm{e}^2\sqrt{r(r-1)} \cdot \max\left(\varepsilon^{d-1}, \mu_{\mathcal{B}}^{d-1}\right) \right) & \text{otherwise.} \end{cases}$$

Proof. This follows from Proposition 2 combined with Lemma 7 after noting that $\mu'_{\mathcal{B}} = 0$ holds. □

Another natural set of bases on which the property $\mu'_{\mathcal{B}} < (r-1)^{-1}$ is satisfied is the random family of sub-Gaussian tensors. Lemma 8 below shows that if all the components of all vectors $\mathbf{y}_k^{(j)}$ (for $j \in [d], k \in [r]$) are normalized independent K-sub-Gaussian random variables (see Definition 5 below), the coherence is actually low with high probability.

DEFINITION 5. *A random variable* ξ *is called* K-sub-Gaussian *if for all* $t \geq 0$,

$$\mathbb{P}\{|\xi| > t\} \leq 2\exp\left(-t^2/K^2\right).$$

Informally, all normal random variables (with any mean and variance), and also those with lighter tails, are K-sub-Gaussian with some proper constant K. All bounded random variables are sub-Gaussian.

LEMMA 8. *Let $\mu > 0$. Let $j \in [d]$ and $\mathcal{Y} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ be a rank-r tensor as per (3.1). Let $n = \min_{i \in [d]} n_i$. If all components of all vectors $\mathbf{y}_k^{(j)}$ are normalized independent mean zero K-sub-Gaussian random variables, with probability at least $1 - 2r^2 d \exp\left(-c\mu^2 n\right)$ the maximum modewise coherence parameter of the tensor $\mathcal{Y}$ is at most μ. Here, c is a positive constant depending only on K.*

Proof. For any $k \in [r]$ and $j \in [d]$ denote $\tilde{\mathbf{y}}_k^{(j)} := \mathbf{y}_k^{(j)} \cdot \|\tilde{\mathbf{y}}_k^{(j)}\|$. By definition, $\tilde{\mathbf{y}}_k^{(j)}$ are independent K-sub-Gaussian random variables for all $k \in [r]$ and $j \in [d]$. Therefore, their norms are of order $\sqrt{n}$ with high probability: for any fixed k, j,

$$\mathbb{P}\left\{n/2 \leq \|\tilde{\mathbf{y}}_k^{(j)}\|_2^2 \leq 2n\right\} \geq 1 - 2\exp\left(-c_1 n/K^4\right)$$

(see, e.g., [58, section 3.1]). Taking union bound, we can conclude that with probability at least $1 - 2rd \exp\left(-c_1 n/K^4\right)$, all vectors $\tilde{\mathbf{y}}_k^{(j)}$ have their norms between $[\sqrt{n/2}, \sqrt{2n}]$.

For any mean zero independent K-sub-Gaussian vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\begin{aligned}
&\mathbb{P}\{|\langle \mathbf{x}, \mathbf{y}\rangle| \geq \mu\|\mathbf{x}\|\|\mathbf{y}\|\} \\
(3.9) \qquad &\leq \mathbb{P}\left\{|\langle \mathbf{x}, \mathbf{y}\rangle| \geq \mu\|\mathbf{y}\|\sqrt{n/2}\right\} + \mathbb{P}\left\{\|\mathbf{x}\| < \sqrt{n/2}\right\}.
\end{aligned}$$

To bound the first term, let us use Hoeffding's inequality (see, e.g., [58, Theorem 2.6.3]). Conditioning on $\mathbf{y}$, we have

$$\mathbb{P}_{\mathbf{x}}\left\{\left|\sum_i x_i y_i\right| \geq \mu\|\mathbf{y}\|\sqrt{n/2}\right\} \leq 2\exp\left(-\frac{c_2 \mu^2 n}{2K^2}\right).$$

Now, let $\tilde{\mathbf{y}}_k^{(j)} = \mathbf{x}$ and $\tilde{\mathbf{y}}_l^{(j)} = \mathbf{y}$. Integrating over $\tilde{\mathbf{y}}_l^{(j)}$ and then taking union bound over all choices of k, l, and j, we get $|\langle \mathbf{y}_k^{(j)}, \mathbf{y}_l^{(j)}\rangle| \leq \mu$ for all component vectors in the tensor $\mathcal{Y}$ with probability at least

$$1 - 2r^2 d \exp\left(-\frac{c_2 \mu^2 n}{2K^2}\right) - 2rd \exp\left(-\frac{c_1 n}{K^4}\right) \geq 1 - 2r^2 d \exp\left(-c\mu^2 n\right).$$

Lemma 8 is proved.   □

The following two elementary corollaries illustrate the applicability of our theory to independent sub-Gaussian tensors. In these corollaries, the term *sub-Gaussian tensor* always refers to a tensor defined as per Lemma 8 and should not be confused with a tensor with sub-Gaussian elements.

COROLLARY 2. *Let $\varepsilon \in (0, 3/4]$. Let $\mathcal{Y}$ be a sub-Gaussian tensor defined as in Lemma 8. For low-rank tensors in high-dimensional spaces, such that*

$$n := \min_{i \in [d]} n_i \geq \frac{\log(r^2 d)}{\varepsilon^2 c}$$

(the small constant c is the same as in Lemma 8), with probability at least $1 - \exp\left(c'\varepsilon^2 n\right)$, Proposition 2 holds with better dependence on ε, namely,

$$\left|\|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2\right| \leq \left(\varepsilon^d r + \varepsilon\right) e^2 \|\boldsymbol{\alpha}\|_2^2.$$

Here, $c' > 0$ is an absolute constant.

Proof. Apply Lemma 8 with $\mu = \varepsilon$. □

COROLLARY 3. *Let $\mathcal{Y}$ be a sub-Gaussian tensor defined as in Lemma 8. If*

$$n := \min_{i \in [d]} n_i \geq C r^{2/d} \log\left(\max\left(r, d\right)\right),$$

with probability at least $1 - \exp\left(-c'n/r^{2/d}\right)$, Lemma 7 gives a nontrivial lower bound $\|\mathcal{Y}\| \geq 0.99\|\boldsymbol{\alpha}\|$. Here, $c' > 0$ is an absolute constant.
In particular, the claim holds when $r \leq C_1^d$ and $n \geq C_2 \max\{r, d\}$.

Proof. Apply Lemma 8 with $\mu = \left(\frac{0.01}{r-1}\right)^{d^{-1}}$. □

Remark 4. Note that in the general case, when r can be as large as $O(n^d)$, the $\mu_{\mathcal{Y}}$ estimate given in Lemma 8 is not strong enough. Indeed, to have a nontrivial probability estimate, one must take $\mu > \sqrt{2d \log n/n}$). However, $\mu_{\mathcal{Y}} \sim \sqrt{d \log n/n}$, together with $r \sim n^d$, do not satisfy the condition of Lemma 7, since $(r-1)\mu_{\mathcal{Y}}^d \sim (dn \log n)^{d/2} \gg 1$.

One could use alternative and more sophisticated anticoncentration results instead of Lemma 7. For example, it was shown recently in [59] that for any $r \leq 0.99n^d$ and under some mild conditions, $\|\mathcal{Y}\| \geq cn^{-d/2}\|\boldsymbol{\alpha}\|_2$ (in the independent sub-Gaussian setting as discussed above). Note that this result contains additional nonfavorable dependence on n. To the best of our knowledge, it is an open question whether general systems of independent (sub-)Gaussian vectors form tensors that satisfy norm anticoncentration such as that in Lemma 7. See also the discussion in [59].

4. Applications to least squares problems and fitting CP models. Now, let us consider the following *fitting* problem. Given tensor $\mathcal{X}$, which is suspected to have (approximately) low CP-rank r, we would like to find the rank-r tensor $\mathcal{Y}$ in the standard form, as per (3.1), being closest to $\mathcal{X}$ in the tensor Euclidean norm. Although the r-dimensional basis (subspace) of $\mathcal{Y}$ is naturally unknown, a common way to tackle the fitting problem is to start with a randomly generated basis, and then update the basis tensors mode by mode, improving the least square error. This brings us to a framework considered in the previous section: a tensor $\mathcal{Y}$ being in some fixed low-dimensional subspace at each step. Since this subspace is changing throughout the fitting process, the oblivious subspace dimension reduction technique is desirable. The fitting problem can be considered as a generalization of the embedding problem introduced in the previous section (with the addition of a potentially full-rank tensor $\mathcal{X}$ that is being approximated).

In this section, we formalize the fitting problem and explain how we propose to use modewise dimension reduction for it. Then, we develop the machinery generalizing our methods from section 3 to incorporate an unknown tensor $\mathcal{X}$. Finally, we propose a more sophisticated two-step dimension reduction process that further improves the resulting dimension for both embedding and fitting problems to almost log-optimal order $\mathcal{O}(r\varepsilon^{-2})$.

As explained above, the common alternating least squares approach for fitting a low-rank CP decomposition along the lines of (3.1) to an arbitrary tensor $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ involves solving a sequence of least squares problems

$$(4.1) \qquad \underset{\tilde{\mathbf{y}}_1^{(j)}, \ldots, \tilde{\mathbf{y}}_r^{(j)} \in \mathbb{C}^{n_j}}{\arg\min} \left\| \mathcal{X} - \sum_{k=1}^{r} \alpha_k \bigcirc_{\ell=1}^{d} \mathbf{y}_k^{(\ell)} \right\|$$

for each $j \in [d]$ after fixing $\{\mathbf{y}_k^{(\ell)}\}_{k \in [r], \ell \in [d] \setminus \{j\}}$. Here, $\mathbf{y}_k^{(j)} = \tilde{\mathbf{y}}_k^{(j)}/\|\tilde{\mathbf{y}}_k^{(j)}\|_2$ for all j, k

and $\alpha_k = \prod_{\ell=1}^d \|\tilde{\mathbf{y}}_k^{(\ell)}\|_2$. One then varies j through all values in $[d]$ computing (4.1) for each j in order to update $\mathbf{y}_k^{(j)}$ for all j, k (potentially cycling through all d modes many times). This makes it particularly important to solve each least squares problem (4.1) efficiently.

Fix $j \in [d]$, and let $\mathbf{e}_h \in \mathbb{C}^{n_j}$ be the hth column of the $n_j \times n_j$ identity matrix. To see how our modewise tensor subspace embeddings can be of value for solving (4.1), one can begin by noting that

$$
\left\| \mathcal{X} - \sum_{k=1}^r \alpha_k \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \right\|^2 = \left\| \mathbf{X}_{(j)} - \sum_{k=1}^r \alpha_k \mathbf{y}_k^{(j)} \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)^\top \right\|_F^2
$$

$$
= \left\| \sum_{h=1}^{n_j} \left(\mathbf{X}_{(j)}^{(h)} - \sum_{k=1}^r \alpha_k y_{k,h}^{(j)} \mathbf{e}_h \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)^\top \right) \right\|_F^2,
$$

where $\mathbf{X}_{(j)}$ denotes mode-j matricization of $\mathcal{X}$, and all the rows of $\mathbf{X}_{(j)}^{(h)} \in \mathbb{C}^{n_j \times \prod_{\ell \neq j} n_\ell}$ are zero except for its hth row, which matches that of $\mathbf{X}_{(j)}$. We may now compute the squared Frobenius norm directly above rowwise and get that

$$
\left\| \mathcal{X} - \sum_{k=1}^r \alpha_k \bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \right\|^2 = \sum_{h=1}^{n_j} \left\| \mathbf{x}_{j,h} - \sum_{k=1}^r \alpha_k y_{k,h}^{(j)} \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right) \right\|_F^2
$$

$$
= \sum_{h=1}^{n_j} \left\| \mathcal{X}^{(j,h)} - \sum_{k=1}^r \alpha_k y_{k,h}^{(j)} \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right\|^2,
$$

where $\mathbf{x}_{j,h} \in \mathbb{C}^{\prod_{\ell \neq j} n_\ell}$ denotes the hth row of $\mathbf{X}_{(j)}$, and $\mathcal{X}^{(j,h)}$ denotes its tensorized version. As a consequence, (4.1) can be decoupled into n_j separate least squares problems of the form

$$
\text{(4.2)} \qquad \underset{\boldsymbol{\alpha}'_{j,h} \in \mathbb{C}^r}{\arg\min} \left\| \mathcal{X}^{(j,h)} - \sum_{k=1}^r \alpha'_{j,h,k} \bigcirc_{\ell \neq j}^d \mathbf{y}_k^{(\ell)} \right\|,
$$

each involving one $(d-1)$-mode mode-j slice, $\mathcal{X}^{(j,h)}$, of the original tensor $\mathcal{X}$.[4] Here $\alpha'_{j,h,k} := \alpha_k y_{k,h}^{(j)}$, where α_k is known for all $k \in [r]$ from (4.1). Note also that these n_j separate least squares problems can, if desired, be solved in parallel for each different $h \in [n_j]$.

In order to solve each least squares problem (4.2) we can now utilize modewise JL embeddings and instead solve the smaller least squares problem

$$
\text{(4.3)} \qquad \underset{\boldsymbol{\alpha}'_{j,h} \in \mathbb{C}^r}{\arg\min} \left\| \mathcal{X}^{(j,h)} \underset{\ell \neq j}{\times} \mathbf{A}_\ell - \sum_{k=1}^r \alpha'_{j,h,k} \bigcirc_{\ell \neq j}^d \mathbf{y}_k^{(\ell)} \underset{\ell \neq j}{\times} \mathbf{A}_\ell \right\|
$$

provided that the $\{\mathbf{y}_k^{(\ell)}\}_{k \in [r]}$ are sufficiently incoherent for all $\ell \in [d] \setminus \{j\}$ (an easy-to-check condition). We can then update each entry of $\tilde{\mathbf{y}}_k^{(j)}$ by setting $\tilde{y}_{k,h}^{(j)} = \alpha'_{j,h,k} / \alpha_k$ for all $h \in [n_j]$ and $k \in [r]$.

---

[4] $\mathcal{X}^{(j,h)}$ is, in fact, the hth mode-j slice of $\mathcal{X}$.

4.1. General modewise JL embeddings for tensors with low modewise coherence. We prove that the method described above works in Theorem 4 below, showing that the solution to (4.3) will be close to that of (4.2) in terms of quality if the matrices $\mathbf{A}_j$ are chosen from appropriate η-optimal JL families of distributions.

THEOREM 4. *Fix $\varepsilon, \eta \in (0, 1/2)$ and $d \geq 3$. Let $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, $n := \max_j n_j \geq 4r + 1$, and let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ spanned by a basis $\mathcal{B} := \{\bigcirc_{\ell=1}^d \mathbf{y}_k^{(\ell)} \mid k \in [r]\}$ of rank-1 tensors, with modewise coherence satisfying $\mu_{\mathcal{B}}^{d-1} < 1/2r$. For each $j \in [d]$ draw $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ with*

(4.4)
$$m_j \geq C_j \cdot rd^3/\varepsilon^2 \cdot \ln\left(n/\sqrt[d]{\eta}\right)$$

from an $(\eta/4d)$-optimal family of JL embedding distributions, where $C_j \in \mathbb{R}^+$ is an absolute constant. Furthermore, let $\mathbf{A} \in \mathbb{C}^{m' \times \prod_{\ell=1}^d m_\ell}$ with

$$m' \geq C'r \cdot \varepsilon^{-2} \cdot \ln\left(\frac{47}{\varepsilon \sqrt[r]{\eta}}\right)$$

be drawn from an $(\eta/2)$-optimal family of JL embedding distributions, where $C' \in \mathbb{R}^+$ is an absolute constant. Define $\tilde{L} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m_1 \times \cdots \times m_d}$ by $L(\mathcal{Z}) = \mathcal{Z} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d$. Then, with probability at least $1 - \eta$, the linear operator $\mathbf{A} \circ \text{vect} \circ \tilde{L} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m'}$ satisfies

$$\left| \left\| \mathbf{A}\left(\text{vect} \circ \tilde{L}\left(\mathcal{X} - \mathcal{Y}\right)\right) \right\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right| \leq \varepsilon \|\mathcal{X} - \mathcal{Y}\|^2$$

for all $\mathcal{Y} \in \mathcal{L}$.

Remark 5 (about r and ε dependence). Fix $d, n,$ and η. Looking at Theorem 4 we can see that its intermediate embedding dimension is

$$\prod_{\ell=1}^d m_\ell \leq C_{d,\eta,n}^d r^d \varepsilon^{-2d},$$

which effectively determines its overall storage complexity. Hence, Theorem 4 will only result in an improved memory complexity over the straightforward single-stage vectorization approach if, e.g., the rank r of $\mathcal{L}$ is relatively small. The purpose of facultative vectorization and subsequent multiplication by an additional JL transform $\mathbf{A}$ in Theorem 4 is to reduce the resulting final embedding dimension to the near-optimal order $\mathcal{O}(r/\varepsilon^2)$ from total dimension $\mathcal{O}_{\eta,n}(d^{3d}r^d\varepsilon^{-2d})$ that we have after the modewise compression.

In order to prove Theorem 4, we first establish that $\|\mathcal{X}^{(j,h)} \times_{\ell \neq j} \mathbf{A}_\ell\| \approx \|\mathcal{X}^{(j,h)}\|$ can also hold for all $j \in [d]$ and $h \in [n_j]$. This is shown in the following lemma which is proven in Appendix B.

LEMMA 9. *Let $\varepsilon \in (0, 1)$, let $\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(p)} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, and let $\mathbf{A}_1 \in \mathbb{C}^{m_1 \times n_1}$ be an $(\varepsilon/\text{e}d)$-JL embedding of all $p(\prod_{\ell=2}^d n_\ell)$ mode-1 fibers of all p of these tensors,*

$$\mathcal{S}_1 := \bigcup_{t \in [p]} \left\{ \mathcal{Z}_{:,i_2,\ldots,i_d}^{(t)} \mid \text{for all } i_\ell \in [n_\ell], \ \ell \in [d] \setminus \{1\} \right\} \subset \mathbb{C}^{n_1}$$

into $\mathbb{C}^{m_1}$. Next, set $\mathcal{Z}^{(1,t)} := \mathcal{Z}^{(t)} \times_1 \mathbf{A}_1 \in \mathbb{C}^{m_1 \times n_2 \times \cdots \times n_d}$ for all $t \in [p]$, and then let $\mathbf{A}_2 \in \mathbb{C}^{m_2 \times n_2}$ be an $(\varepsilon/\text{e}d)$-JL embedding of all $p(m_1 \prod_{\ell=3}^d n_\ell)$ mode-2 fibers

$$\mathcal{S}_2 := \bigcup_{t \in [p]} \left\{ \mathcal{Z}_{i_1,:,i_3,\ldots,i_d}^{(1,t)} \mid \text{for all } i_1 \in [m_1] \text{ and } i_\ell \in [n_\ell], \ \ell \in [d] \setminus [2] \right\} \subset \mathbb{C}^{n_2}$$

into $\mathbb{C}^{m_2}$. *Continuing inductively, for each* $j \in [d] \setminus [2]$ *and* $t \in [p]$, *set* $\mathcal{Z}^{(j-1,t)} := \mathcal{Z}^{(j-2,t)} \times_{j-1} \mathbf{A}_{j-1} \in \mathbb{C}^{m_1 \times \cdots \times m_{j-1} \times n_j \times \cdots \times n_d}$, *and then let* $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ *be an* $(\varepsilon/\text{e}d)$-*JL embedding of all* $p(\prod_{\ell=1}^{j-1} m_\ell)(\prod_{\ell=j+1}^{d} n_\ell)$ *mode-*j *fibers*

$$\mathcal{S}_j := \bigcup_{t \in [p]} \left\{ \mathcal{Z}^{(j-1,t)}_{i_1, \ldots, i_{j-1}, :, i_{j+1}, \ldots, i_d} \mid \text{for all } i_\ell \in [m_\ell], \ell \in [j-1] \text{ and } i_\ell \in [n_\ell], \ell \in [d] \setminus [j] \right\} \subset \mathbb{C}^{n_j}$$

into $\mathbb{C}^{m_j}$. *Then,*

$$\left| \left\| \mathcal{Z}^{(t)} \right\|^2 - \left\| \mathcal{Z}^{(t)} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d \right\|^2 \right| \leq \varepsilon \left\| \mathcal{Z}^{(t)} \right\|^2$$

will hold for all $t \in [p]$.

With Lemma 9 in hand we can now prove that the solution to (4.3) will be close to that of (4.2) in terms of quality if the matrices $\mathbf{A}_j$ are chosen appropriately. We have the following general result which directly applies to least squares problems as per (4.3) when $\tilde{L}(\mathcal{Z}) := \mathcal{Z} \times_{\ell \neq j} \mathbf{A}_\ell$ and $\mathbf{A} = \mathbf{I}$.

THEOREM 5 (embeddings for compressed least squares). *Let* $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, *let* $\mathcal{L}$ *be an* r-*dimensional subspace of* $\mathbb{C}^{n_1 \times \cdots \times n_d}$ *spanned by a set of orthonormal basis tensors* $\{\mathcal{T}_k\}_{k \in [r]}$, *and let* $\mathbb{P}_{\mathcal{L}^\perp} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{n_1 \times \cdots \times n_d}$ *be the orthogonal projection operator on the orthogonal complement of* $\mathcal{L}$. *Fix* $\varepsilon \in (0, 1)$, *and suppose that the linear operator* $\tilde{L} : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^{m_1 \times \cdots \times m_{d'}}$ *has both of the following properties:*
  (i) $\tilde{L}$ *is an* $(\varepsilon/6)$-*JL embedding of all* $\mathcal{Y} \in \mathcal{L} \cup \{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\}$ *into* $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$, *and*
  (ii) $\tilde{L}$ *is an* $(\varepsilon/24\sqrt{r})$-*JL embedding of the* $4r$ *tensors*

$$\mathcal{S}' := \bigcup_{k \in [r]} \left\{ \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} - \mathcal{T}_k, \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} + \mathcal{T}_k, \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} - \mathrm{i}\mathcal{T}_k, \right.$$
$$\left. \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} + \mathrm{i}\mathcal{T}_k \right\} \subset \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$$

into $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$.
Furthermore, let $\mathrm{vect} : \mathbb{C}^{m_1 \times \cdots \times m_{d'}} \to \mathbb{C}^{\prod_{\ell=1}^{d'} m_\ell}$ *be a reshaping vectorization operator, and let* $\mathbf{A} \in \mathbb{C}^{m \times \prod_{\ell=1}^{d'} m_\ell}$ *be an* $(\varepsilon/3)$-*JL embedding of the* $(r+1)$-*dimensional subspace*

$$\mathcal{L}' := \mathrm{span}\left\{ \mathrm{vect} \circ \tilde{L}(\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})), \ \mathrm{vect} \circ \tilde{L}(\mathcal{T}_1), \ \ldots, \ \mathrm{vect} \circ \tilde{L}(\mathcal{T}_r) \right\} \subset \mathbb{C}^{\prod_{\ell=1}^{d'} m_\ell}$$

into $\mathbb{C}^m$. *Then,*

$$\left| \left\| \mathbf{A} \left(\mathrm{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y}) \right) \right\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right| \leq \varepsilon \|\mathcal{X} - \mathcal{Y}\|^2$$

holds for all $\mathcal{Y} \in \mathcal{L}$.

Proof. Note that the theorem will be proven if $\tilde{L}$ is an $(\varepsilon/3)$-JL embedding of all tensors of the form $\{\mathcal{X} - \mathcal{Y} \mid \mathcal{Y} \in \mathcal{L}\}$ into $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$ since any such tensor $\mathcal{X} - \mathcal{Y}$

will also have $\text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y}) \in \mathcal{L}'$ so that

$$
\left| \left\| \mathbf{A}\left(\text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y})\right) \right\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right|
$$

$$
\leq \left| \left\| \mathbf{A}\left(\text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y})\right) \right\|_2^2 - \left\| \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|^2 \right| + \left| \left\| \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right|
$$

$$
\leq \left| \left\| \mathbf{A}\left(\text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y})\right) \right\|_2^2 - \left\| \text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|_2^2 \right| + \frac{\varepsilon}{3}\|\mathcal{X} - \mathcal{Y}\|^2
$$

$$
\leq \frac{\varepsilon}{3}\left\| \text{vect} \circ \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|_2^2 + \frac{\varepsilon}{3}\|\mathcal{X} - \mathcal{Y}\|^2
$$

$$
= \frac{\varepsilon}{3}\left\| \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|^2 + \frac{\varepsilon}{3}\|\mathcal{X} - \mathcal{Y}\|^2
$$

$$
\leq \frac{\varepsilon}{3}\left(1 + \frac{\varepsilon}{3}\right)\|\mathcal{X} - \mathcal{Y}\|^2 + \frac{\varepsilon}{3}\|\mathcal{X} - \mathcal{Y}\|^2 \ \leq \ \varepsilon\|\mathcal{X} - \mathcal{Y}\|^2.
$$

Let $\mathbb{P}_{\mathcal{L}}$ be the orthogonal projection operator onto $\mathcal{L}$. Our first step in establishing that $\tilde{L}$ is an $(\varepsilon/3)$-JL embedding of all tensors of the form $\{\mathcal{X} - \mathcal{Y} \mid \mathcal{Y} \in \mathcal{L}\}$ into $\mathbb{C}^{m_1 \times \cdots \times m_{d'}}$ will be to show that $\tilde{L}$ preserves all the angles between $\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})$ and $\mathcal{L}$ well enough that the Pythagorean theorem

$$
\|\mathcal{X} - \mathcal{Y}\|^2 \ = \ \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X}) + \mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2 \ = \ \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|^2 + \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2
$$

still approximately holds for all $\mathcal{Y} \in \mathcal{L}$ after $\tilde{L}$ is applied. Toward that end, let $\boldsymbol{\gamma} \in \mathbb{C}^r$ be such that $\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y} = \sum_{k \in [r]} \gamma_k \mathcal{T}_k$, and note that $\|\boldsymbol{\gamma}\|_2 = \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|$ due to the orthonormality of $\{\mathcal{T}_k\}_{k \in [r]}$. Appealing to Lemma 2 we now have that

(4.5)

$$
\left| \left\langle \tilde{L}(\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}),\ \tilde{L}(\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})) \right\rangle \right| = \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\| \left| \sum_{k \in [r]} \gamma_k \left\langle \tilde{L}(\mathcal{T}_k),\ \tilde{L}\left(\frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|}\right) \right\rangle \right|
$$

$$
\leq \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\| \left(\frac{\varepsilon}{6\sqrt{r}}\right) \sum_{k \in [r]} |\gamma_k| \leq \frac{\varepsilon}{6}\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|\,\|\boldsymbol{\gamma}\|_2
$$

$$
\leq \frac{\varepsilon}{12}\left(\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|^2 + \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2\right) = \frac{\varepsilon}{12}\|\mathcal{X} - \mathcal{Y}\|^2.
$$

Using (4.5) we can now see that

$$
\left| \left\| \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|_2^2 - \|\mathcal{X} - \mathcal{Y}\|^2 \right|
$$

$$
= \left| \left\| \tilde{L}(\mathcal{X} - \mathcal{Y}) \right\|_2^2 - \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|^2 - \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2 \right|
$$

$$
\leq \left| \left\| \tilde{L}(\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})) \right\|^2 - \|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|^2 \right| + \left| \left\| \tilde{L}(\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}) \right\|^2 - \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2 \right|
$$

$$
+ 2\left| \left\langle \tilde{L}(\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}),\ \tilde{L}(\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})) \right\rangle \right|
$$

$$
\leq \frac{\varepsilon}{6}\left(\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|^2 + \|\mathbb{P}_{\mathcal{L}}(\mathcal{X}) - \mathcal{Y}\|^2 + \|\mathcal{X} - \mathcal{Y}\|^2\right) \ = \ \frac{\varepsilon}{3}\|\mathcal{X} - \mathcal{Y}\|^2.
$$

Thus, $\tilde{L}$ has the desired JL embedding property required to conclude the proof. $\square$

Theorems 2 and 5 together with Lemma 9 can now be used to demonstrate the existence of a large range of modewise Johnson–Lindenstrauss transforms (JLTs) for oblivious tensor subspace embeddings. The following modewise JLT result for tensors describes the compression one can achieve from Theorem 5 if the linear operator L one employs is formed using j-mode products (as considered in Proposition 2) with $\mathbf{A}_j \in \mathbb{C}^{m_j \times n_j}$ taken from η-optimal families of JL embedding distributions (in the sense of Definition 2).

We are now ready to complete the proof of Theorem 4.

Proof of Theorem 4. To begin, we note that $\mathbf{A}$ will satisfy the conditions required by Theorem 5 with probability at least $1 - \eta/2$ as a consequence of Lemma 3. Thus, if we can also establish that $\tilde{L}$ will satisfy the conditions required by Theorem 5 with probability at least $1 - \eta/2$, we will be finished with our proof by Theorem 5 and the union bound.

To establish that $\tilde{L}$ satisfies the conditions required by Theorem 5 with probability at least $1 - \eta/2$, it suffices to prove that

(a) $\tilde{L}$ will be an $(\varepsilon/6)$-JL embedding of all $\mathcal{Y} \in \mathcal{L}$ into $\mathbb{C}^{m_1 \times \cdots \times m_d}$ with probability at least $1 - \eta/4$, and that

(b) $\tilde{L}$ will be an $(\varepsilon/24\sqrt{r})$-JL embedding of the $4r + 1$ tensors $\mathcal{S}' \cup \{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\} \subset \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ into $\mathbb{C}^{m_1 \times \cdots \times m_d}$ with probability at least $1 - \eta/4$, where the set $\mathcal{S}'$ is defined as in Theorem 5,

and apply yet another union bound.

To show that (a) holds, we will utilize Proposition 2 and Lemma 7. Since each $\mathbf{A}_j$ matrix is an $(\eta/4d)$-optimal JL embedding and the sets $\mathcal{S}'_j$ (defined as in Proposition 2) are such that $|\mathcal{S}'_j| < n^d$, we know that each $\mathbf{A}_j$ is an $(\varepsilon/480d\sqrt{r})$-JL embedding of $\mathcal{S}'_j$ into $\mathbb{C}^{m_j}$ with probability[5] at least $1 - \eta/4d$. Thus, Proposition 2 holds with $\varepsilon \to \varepsilon/120\sqrt{r}$ with probability at least $1 - \eta/4$. Note that the modewise coherence assumption that $\mu_{\mathcal{B}}^{d-1} < 1/2r$ both allows ε^{d-1} to reduce the $\sqrt{r(r-1)}$ factor in (3.7) to a size less than one for any $\varepsilon \le 1/\sqrt{r} \le (1/r)^{1/(d-1)}$, and allows Lemma 7 to guarantee that $\|\boldsymbol{\alpha}\|_2^2 < 2 \|\mathcal{Y}\|^2$ holds for all $\mathcal{Y} \in \mathcal{L}$. Hence, applying Proposition 2 with $\varepsilon \to \varepsilon/120\sqrt{r}$ will ensure that $\tilde{L}$ is an $(\varepsilon/6)$-JL embedding of all $\mathcal{Y} \in \mathcal{L}$ into $\mathbb{C}^{m_1 \times \cdots \times m_d}$.

To show that (b) holds we will utilize Lemma 9. Note that the $\mathcal{S}_j$ sets defined in Lemma 9 all have cardinalities $|\mathcal{S}_j| \le p \cdot n^{d-1}$, where $p = 4r + 1 \le n$ in our current setting. As a consequence we can see that the conditions of Lemma 9 will be satisfied with $\varepsilon \to \varepsilon/24\sqrt{r}$ for all $j \in [d]$ with probability at least $1 - \eta/4$ by the union bound. Hence, both (a) and (b) hold and our proof is concluded. $\square$

We will now consider a final tensor subspace embedding result concerning a special case of modewise JL embeddings that is also made possible by our work above. This result will exhibit better dependence with respect to both ε and r than what is achieved by the more general modewise embedding constructions in Theorem 4.

4.2. Fast and memory efficient modewise JL embeddings for tensors. In this section we consider a fast JLT for tensors recently introduced in [32], which is effectively based on applying fast JLTs [37] in a modewise fashion.[6] Given a tensor

---

[5] Here we also implicitly use the fact that $\sqrt[d]{d} \le \sqrt[\varepsilon]{e}$ holds for all $d > 0$ in order to avoid a $\sqrt[d]{d}$ term appearing inside the logarithm in (4.4).

[6] In fact, the fast transform described here differs cosmetically from the form in which it is presented in [32]. However, one can easily see they are equivalent using (2.8).

$\mathcal{Z} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$ the transform takes the form

$$(4.6) \qquad L_{\mathrm{FJL}}(\mathcal{Z}) := \sqrt{\frac{N}{m}} \, \mathbf{R} \left(\mathrm{vect}\left(\mathcal{Z} \times_1 \mathbf{F}_1 \mathbf{D}_1 \cdots \times_d \mathbf{F}_d \mathbf{D}_d \right) \right),$$

where $\mathrm{vect} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^N$ for $N := \prod_{\ell=1}^d n_\ell$ is the vectorization operator, $\mathbf{R} \in \{0, 1\}^{m \times N}$ is a matrix containing m rows selected randomly from the $N \times N$ identity matrix, $\mathbf{F}_\ell \in \mathbb{C}^{n_\ell \times n_\ell}$ is a unitary discrete Fourier transform matrix for all $\ell \in [d]$, and $\mathbf{D}_\ell \in \mathbb{C}^{n_\ell \times n_\ell}$ is a diagonal matrix with n_ℓ random ± 1 entries for all $\ell \in [d]$. The following theorem is proven about this transform in [32, 37].

THEOREM 6 (see Theorem 2.1 and Remark 4 in [32]). *Fix $d \geq 1$, $\varepsilon, \eta \in (0, 1)$, and $N \geq C'/\eta$ for a sufficiently large absolute constant $C' \in \mathbb{R}^+$. Consider a finite set $\mathcal{S} \subset \mathbb{C}^{n_1 \times \cdots \times n_d}$ of cardinality $p = |\mathcal{S}|$, and let $L_{\mathrm{FJL}} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^m$ be defined as above in* (4.6) *with*

$$m \; \geq \; C \left[\varepsilon^{-2} \cdot \log^{2d-1} \left(\frac{\max(p, N)}{\eta} \right) \cdot \log^4 \left(\frac{\log\left(\frac{\max(p, N)}{\eta} \right)}{\varepsilon} \right) \cdot \log N \right],$$

where $C > 0$ is an absolute constant. Then with probability at least $1 - \eta$ the linear operator L_{FJL} is an ε-JL embedding of $\mathcal{S}$ into $\mathbb{C}^m$. If $d = 1$, then we may replace $\max(p, N)$ with p inside all of the logarithmic factors above (see [37]).

Note that the fast transform L_{FJL} requires only $\mathcal{O}\left(m \log N + \sum_\ell n_\ell\right)$ i.i.d. random bits and memory for storage. Thus, it can be used to produce fast and low memory complexity oblivious subspace embeddings. The next theorem does so.

THEOREM 7. *Fix $\varepsilon, \eta \in (0, 1/2)$ and $d \geq 2$. Let $\mathcal{X} \in \mathbb{C}^{n_1 \times \cdots \times n_d}$, let $N = \prod_{\ell=1}^d n_\ell \geq 4C'/\eta$ for an absolute constant $C' > 0$, let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{C}^{n_1 \times \cdots \times n_d}$ for $\max\left(2r^2 - r, 4r\right) \leq N$, and let $L_{\mathrm{FJL}} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m_1}$ be defined as above in* (4.6) *with*

$$m_1 \; \geq \; C_1 \left[C_2^d \left(\frac{r}{\varepsilon} \right)^2 \cdot \log^{2d-1} \left(\frac{N}{\eta} \right) \cdot \log^4 \left(\frac{\log\left(\frac{N}{\eta} \right)}{\varepsilon} \right) \cdot \log N \right],$$

where $C_1, C_2 > 0$ are absolute constants. Furthermore, let $\mathbf{L}'_{\mathrm{FJL}} \in \mathbb{C}^{m_2 \times m_1}$ be defined as above in (4.6) *for $d = 1$ with*

$$m_2 \; \geq \; C_3 \left[r \cdot \varepsilon^{-2} \cdot \log \left(\frac{47}{\varepsilon \sqrt[r]{\eta}} \right) \cdot \log^4 \left(\frac{r \log\left(\frac{47}{\varepsilon \sqrt[r]{\eta}} \right)}{\varepsilon} \right) \cdot \log m_1 \right],$$

where $C_3 > 0$ is an absolute constant. Then, with probability at least $1 - \eta$ it will be the case that

$$\left| \left\| \mathbf{L}'_{\mathrm{FJL}} \left(L_{\mathrm{FJL}} (\mathcal{X} - \mathcal{Y}) \right) \right\|_2^2 - \| \mathcal{X} - \mathcal{Y} \|^2 \right| \leq \varepsilon \, \| \mathcal{X} - \mathcal{Y} \|^2$$

holds for all $\mathcal{Y} \in \mathcal{L}$.

In addition, the $(\mathbf{L}'_{\mathrm{FJL}}, L_{\mathrm{FJL}})$ transform pair requires only $\mathcal{O}\left(m_1 \log N + \sum_\ell n_\ell\right)$ random bits and memory for storage (assuming without loss of generality that $m_2 \leq m_1$), and $\mathbf{L}'_{\mathrm{FJL}} \circ L_{\mathrm{FJL}} : \mathbb{C}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^{m_2}$ can be applied to any tensor in just $\mathcal{O}\left(N \log N\right)$-time.

Proof. Let $\{\mathcal{T}_k\}_{k\in[r]}$ be an orthonormal basis for $\mathcal{L}$ (note that these basis tensors need not be low-rank), and let $\mathbb{P}_{\mathcal{L}^\perp} : \mathbb{C}^{n_1\times\cdots\times n_d} \to \mathbb{C}^{n_1\times\cdots\times n_d}$ be the orthogonal projection operator onto the orthogonal complement of $\mathcal{L}$. Theorem 5 combined with Lemmas 4 and 3 implies that the result will be proven if all of the following hold:

(i) L_{FJL} is an $(\varepsilon/24r)$-JL embedding of the $2r^2 - r$ tensors

$$\left(\bigcup_{1\le h<k\le r} \{\mathcal{T}_k - \mathcal{T}_h, \mathcal{T}_k + \mathcal{T}_h, \mathcal{T}_k - \mathrm{i}\mathcal{T}_h, \mathcal{T}_k + \mathrm{i}\mathcal{T}_h\}\right) \bigcup \{\mathcal{T}_k\}_{k\in[r]} \subset \mathcal{L}$$

into $\mathbb{C}^{m_1}$,

(ii) L_{FJL} is an $(\varepsilon/6)$-JL embedding of $\{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\}$ into $\mathbb{C}^{m_1}$,

(iii) L_{FJL} is an $(\varepsilon/24\sqrt{r})$-JL embedding of the $4r$ tensors

$$\bigcup_{k\in[r]} \left\{ \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} - \mathcal{T}_k, \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} + \mathcal{T}_k, \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} - \mathrm{i}\mathcal{T}_k, \right.$$
$$\left. \frac{\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})}{\|\mathbb{P}_{\mathcal{L}^\perp}(\mathcal{X})\|} + \mathrm{i}\mathcal{T}_k \right\} \subset \mathbb{C}^{n_1\times\cdots\times n_d}$$

into $\mathbb{C}^{m_1}$, and

(iv) $\mathbf{L}'_{\mathrm{FJL}}$ is an $(\varepsilon/6)$-JL embedding of a minimal $(\varepsilon/16)$-cover, $\mathcal{C}$, of the r-dimensional Euclidean unit sphere in the subspace $\mathcal{L}' \subset \mathbb{C}^{m_1}$ from Theorem 5 with $L = L_{\mathrm{FJL}}$ into $\mathbb{C}^{m_2}$. Here we note that $|\mathcal{C}| \le \left(\frac{47}{\varepsilon}\right)^r$.

Furthermore, if m_1 and m_2 are chosen as above for sufficiently large absolute constants C_1, C_2, and C_3, then Theorem 6 implies that each of (i)–(iv) above will fail to hold with probability at most $\eta/4$. The desired result now follows from the union bound.

The number of random bits and storage complexity follow directly from Theorem 6 after noting that each row of $\mathbf{R}$ in (4.6) is determined by $\mathcal{O}(\log N)$ bits. The fact that $\mathbf{L}'_{\mathrm{FJL}} \circ L_{\mathrm{FJL}}$ can be applied to any tensor $\mathcal{Z}$ in $\mathcal{O}(N\log N)$-time again follows from the form of (4.6). Note that each j-mode product with $\mathbf{F}_j\mathbf{D}_j$ involves $\prod_{\ell\ne j} n_\ell$ multiplications of $\mathbf{F}_j\mathbf{D}_j$ against all the mode-j fibers of the given tensor $\mathcal{Z}$, each of which can be performed in $\mathcal{O}(n_j\log(n_j))$-time using fast Fourier transform techniques (or approximated even more quickly using sparse Fourier transform techniques if n_j is itself very large; see, e.g., [24, 44, 12, 29, 30, 51]). The required vectorization and applications of $\mathbf{R}$ can then be performed in just $\mathcal{O}(N)$-time thereafter. Finally, Fourier transform techniques can again be used to also apply $\mathbf{L}'_{\mathrm{FJL}}$ in $\mathcal{O}(m_1\log m_1)$-time. $\square$

Remark 6. To recap, in sections 4.1 and 4.2 we presented two different results concerning modewise oblivious JL embeddings for low-rank tensor subspaces, Theorem 7 and Theorem 4. Unlike Theorem 3, both are suited for tensor low-rank fitting applications since they allow for an affine shift of an arbitrary low-rank tensor subspace $\mathcal{L}$ by an arbitrary (and not necessarily low-rank) fixed tensor $\mathcal{X}$.

Fix d, n, N, and η. Recalling Remark 5 we can see that the intermediate embedding dimension provided by Theorem 4 is $\prod_{\ell=1}^d m_\ell \le C_{d,\eta,n}^d r^d \varepsilon^{-2d}$. In comparison we can see that Theorem 7 achieves an intermediate embedding dimension of size

$$m_1 \le C_{d,\eta,N}^d \left(\frac{r}{\varepsilon}\right)^2 \cdot \log^4\left(\frac{C_{d,\eta,N}}{\varepsilon}\right).$$

Hence, Theorem 7 provides a significantly better intermediate embedding dimension for large d (with respect to r and ε dependence) than Theorem 4 does despite the

fact that both theorems ultimately achieve a near-optimal final embedding dimension. Ultimately, this means that Theorem 7 provides more compactly storable multistage JL embeddings when d is large than Theorem 4 does. Additionally, Theorem 7 does not require the basis tensors of any low-rank subspace to which it is applied to all be rank-1 tensors, an advantage which is not employed in the framework of tensor low-rank fitting problems, but which might be useful in other settings.

On the other hand, Theorem 4 is significantly more general for tensor subspaces with rank-1 bases that have low modewise coherence: it guarantees JL embedding properties for modewise products by any matrices from a large class of almost optimal JL embedding matrices including, e.g., sparse JL embedding matrices. In contrast, Theorem 7 relies on a very particular modewise operation based on discrete Fourier transform (DFT) matrices.

We are now prepared to consider the numerical performance of such modewise JL transforms.

5. Experiments. In this section it is shown that the norms of several different types of (approximately) low-rank data can be preserved using JL embeddings, and trial least squares experiments with compressed tensor data are also performed to show the effect of these embeddings on solutions to least squares problems. All experiments were carried out in MATLAB. The data sets used in the experiments consist of the following:

1. *MRI data.* This data set contains three 3-mode MRI images of size $240 \times 240 \times 155$ [1].
2. *Randomly generated data.* This data set contains 10 rank-10 4-mode tensors. Each test tensor is a $100 \times 100 \times 100 \times 100$ tensor that is created by adding 10 randomly generated rank-1 tensors. More specifically, each rank-10 tensor is generated according to

$$\mathcal{X}^{(m)} = \sum_{k=1}^{r} \bigcirc_{j=1}^{d} \mathbf{y}_k^{(j)},$$

where $m \in [10]$, $r = 10$, $d = 4$, and $\mathbf{y}_k^{(j)} \in \mathbb{R}^{100}$. In the Gaussian case, each entry of $\mathbf{y}_k^{(j)}$ is drawn independently from the standard Gaussian distribution $\mathcal{N}(0,1)$. In the case of coherent data, low-variance Gaussian noise is added to a constant, i.e., each entry $\mathbf{y}_{k,\ell}^{(j)}$ of $\mathbf{y}_k^{(j)}$ is set as $1 + \sigma g_{k,\ell}^{(j)}$, with $g_{k,\ell}^{(j)}$ being an i.i.d. standard Gaussian random variable defined above and σ^2 denoting the desired variance. In the experiments of this section, $\sigma = \sqrt{0.1}$ is used. In both cases, the 2-norm of $\mathbf{y}_k^{(j)}$ is also normalized to 1.

The reason for running experiments on both Gaussian and coherent data is to show that although coherence requirements presented in section 3 are used to help get general theoretical results for a large class of modewise JL embeddings, they do not seem to be necessary in practice.

When JL embeddings are applied, experiments are performed using Gaussian JL matrices as well as fast JL matrices. For Gaussian JL, $\mathbf{A}_j = \frac{1}{\sqrt{m}}\mathbf{G}$ is used for all $j \in [d]$, where m is the target dimension, and each entry in $\mathbf{G}$ is an i.i.d. standard Gaussian random variable $\mathbf{G}_{i,j} \sim \mathcal{N}(0,1)$. For fast JL, $\mathbf{A}_j = \frac{1}{\sqrt{m}}\mathbf{RFD}$ is used for all $j \in [d]$, where $\mathbf{R}$ denotes the random restriction matrix, $\mathbf{F}$ is the unitary DFT matrix scaled by $\sqrt{n_j}$,[7] and $\mathbf{D}$ is a diagonal matrix with Rademacher random variables

---

[7]Recall that n_j is the size of the mode-j fibers of the input tensor.

forming its diagonal [37]. The embedded version of a test tensor $\mathcal{X}$ is always denoted by $L(\mathcal{X})$ and is calculated by

$$(5.1) \qquad L(\mathcal{X}) = \begin{cases} \mathcal{X} \times_1 \mathbf{A}_1 \times \cdots \times_d \mathbf{A}_d, & \text{1-stage JL}, \\[2mm] \mathbf{A}\left(\text{vect}\left(\mathcal{X} \times_1 \mathbf{A}_1 \times \cdots \times_d \mathbf{A}_d\right)\right), & \text{2-stage JL}, \end{cases}$$

where $\mathbf{A}$ is a JL matrix used in the second stage. Obviously, $L(\mathcal{X})$ is a vector in the 2-stage case.

5.1. Effect of JL embeddings on norm. In this section, numerical results have been presented showing the effect of modewise JL embedding on the norm of three MRI 3-mode images treated as generic tensors as well as randomly generated data.

The compression ratio for the jth mode, denoted by $c_1^{(j)}$, is defined as the compression in the size of each of the mode-j fibers, i.e.,

$$c_1^{(j)} = \frac{m_j}{n_j}.$$

The target dimension m_j in JL matrices is chosen as $m_j = \lceil c_1 n_j \rceil$ for all $j \in [d]$ to ensure that *at least* a fraction c_1 of the ambient dimension in each mode is preserved. In the experiments, the compression ratio is set to be the same for all modes, i.e., $c_1^{(j)} = c_1$ for all $j \in [d]$. In the case of a 2-stage JL embedding, the target dimension m of the secondary JL embedding is chosen as

$$m = \lceil c_2 N \rceil,$$

where c_2 is the compression ratio in the second stage, and N is the length of the vectorized projected tensor after the modewise JL embedding. The total achieved compression is calculated by $c_{tot} = c_2(\prod_{j=1}^{d} c_1^{(j)})$. When the 2nd-stage embedding is skipped, $c_{tot} = \prod_{j=1}^{d} c_1^{(j)}$. In all experiments of section 5, when a 2-stage embedding is performed, $c_2 = 0.05$. Also, in the figure legends, when two JL types are listed together, the first and second terms refer to the first and second stages, respectively. For example, in "Gaussian+RFD," Gaussian and RFD JL embeddings were used in the first and second stages, respectively. The term "vec" in the legends refers to vectorizing the data.

Assuming $\mathcal{X}$ denotes the original tensor and $L(\mathcal{X})$ is the projected result, the relative norm of $\mathcal{X}$ is defined by

$$c_{n,\mathcal{X}} = \frac{\|L(\mathcal{X})\|}{\|\mathcal{X}\|}.$$

The results of this section depict the interplay between $c_{n,\mathcal{X}}$ and c_1 for randomly generated data, and between $c_{n,\mathcal{X}}$ and c_{tot} for MRI data, where the numbers have been averaged over 1000 trials as well as over all samples for each value of c_1 or c_{tot}. In the case of Figure 2, 1000 randomly generated JL matrices were applied to each mode of all 10 randomly generated tensors. The results there indicate that the modewise embedding methods proposed herein still work on relatively coherent data despite the incoherence assumptions utilized in their theoretical analysis (recall section 3). In Figure 3, 1000 JL embedding choices have been averaged over each of the three MRI images as well as the three images themselves. As expected, it can be observed in
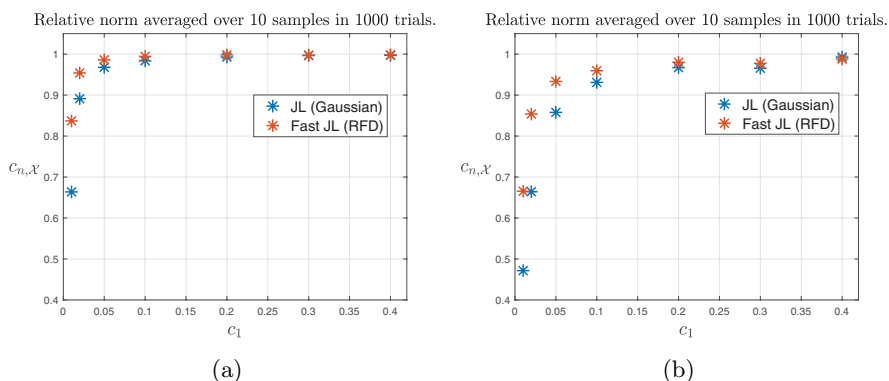
FIG. 2. *Relative norm of randomly generated 4-dimensional data. Here, the total compression will be $c_{tot} = c_1^4$. (a) Gaussian data. (b) Coherent data. Note that the modewise approach still preserves norms well for the coherent data, indicating that the incoherence assumptions utilized in section 3 can likely be relaxed.*

both figures that increasing the compression ratio leads to better norm (and distance) preservation.

The MRI data experiments were done using various combinations of JL matrices in the first and second stages and were compared with the 1-stage (modewise) case and also JL applied to vectorized data. In Figure 3(b), the runtime plots show that vectorizing the data before applying JL embeddings is the most computationally intensive way of compressing the data, although it preserves norms the best, as Figure 3(a) demonstrates. Due to the small mode sizes of the MRI data used in the experiments, modewise fast JL does not outperform modewise Gaussian JL in terms of computational efficiency in the modewise embeddings as one might initially expect (see the red and blue curves). This is likely due to the facts that the individual mode sizes are too small to benefit from the FFT (recall that all modes are ≤ 240 in size) and that the Fourier methods need to use less efficient complex number arithmetic. However, when the 2-stage JL is employed for larger compression ratios, the vectorized data after the first stage compression is large enough to make the efficiency of fast JL over Gaussian JL embeddings clear (compare, e.g., the yellow and purple curves).

5.2. Effect of JL embeddings on least squares solutions. In this section, the first sample of the three MRI data samples is used in the experiments. First, we show that this MRI sample has a relatively low-rank CP representations by plotting its CP reconstruction error for various values of rank. Next, the effect of modewise JL on least squares solutions is investigated by solving for the coefficients of the CP decomposition of the MRI sample in a least squares problem. This will be done by performing 1-stage (modewise) and 2-stage JL on the data, which we call compressed least squares, and will be compared with the case where a regular uncompressed least squares problem is solved instead.

5.2.1. CPD reconstruction. Before the experimental results are shown, a short description of the basic form of CPD calculation is presented as well as how the number of rank-1 tensor, r, is chosen. Given a tensor $\mathcal{X}$, assume r is known beforehand. The problem is now the calculation of $\mathbf{y}_k^{(j)}$ for $j \in [d]$ and $k \in [r]$ and $\boldsymbol{\alpha}$
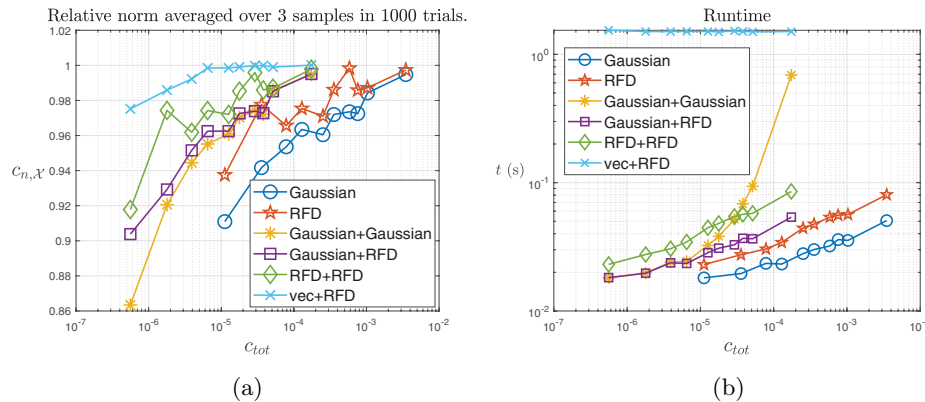
FIG. 3. *Simulation results averaged over* 1,000 *trials for three MRI data samples, where each sample is* 3-*dimensional. In the* 2-*stage cases,* $c_2 = 0.05$ *has been used.* (a) *Relative norm.* (b) *Runtime.*

in (3.1), i.e., the solution to

$$(5.2) \qquad \min_{\hat{\mathcal{X}}} \|\mathcal{X} - \hat{\mathcal{X}}\| \text{ with } \hat{\mathcal{X}} = \sum_{k=1}^{r} \alpha_k \, \mathbf{y}_k^{(1)} \bigcirc \mathbf{y}_k^{(2)} \bigcirc \cdots \bigcirc \mathbf{y}_k^{(d)}.$$

As the Euclidean norm of a d-mode tensor is equal to the Frobenius norm of its mode-j unfoldings for $j \in [d]$, by letting $\mathbf{y}_k^{(j)}$ be the kth column of a matrix $\mathbf{Y}^{(j)} \in \mathbb{C}^{n_j \times r}$, we see that the above minimization problem can be written as

$$\min_{\hat{\mathbf{Y}}^{(j)}} \left\| \mathbf{X}_{(j)} - \hat{\mathbf{Y}}^{(j)} \left(\mathbf{Y}^{(d)} \odot \cdots \odot \mathbf{Y}^{(j+1)} \odot \mathbf{Y}^{(j-1)} \odot \cdots \odot \mathbf{Y}^{(1)} \right)^{\top} \right\|_{\mathrm{F}},$$

where $\hat{\mathbf{Y}}^{(j)} = \mathbf{Y}^{(j)} \mathrm{diag}(\boldsymbol{\alpha})$, and $\odot$ denotes the Khatri–Rao product defined as the columnwise matching Kronecker product. The operator $\mathrm{diag}(\cdot)$ creates a diagonal matrix with $\boldsymbol{\alpha}$ as its diagonal. Once solved for, the columns of $\hat{\mathbf{Y}}^{(j)}$ can then be normalized and used to form the coefficients $\alpha_k = \prod_{j=1}^{d} \|\hat{\mathbf{y}}_k^{(j)}\|_2$ for $k \in [r]$, although this is optional, i.e., if the columns are not normalized, all the coefficients α_k in the factorization will be ones. This procedure is repeated iteratively until the fit ceases to improve (the objective function stops improving with respect to a tolerance) or the maximum number of iterations is exhausted. This procedure is known as CPD-ALS[8] [36]. To choose the rank of the decomposition as well as obtain the best estimates for $\mathbf{Y}^{(j)}$, a commonly used consistency diagnostic called CORCONDIA[9] can be employed [13].

In the remainder of this section, the relative reconstruction error of CPD is calculated and plotted for various values of rank r. Assuming $\mathcal{X}$ represents the data, this error is defined as

$$e_{cpd} = \frac{\|\mathcal{X} - \hat{\mathcal{X}}\|}{\|\mathcal{X}\|},$$

where $\hat{\mathcal{X}}$ denotes the reconstruction of $\mathcal{X}$. Figure 4 displays the results.

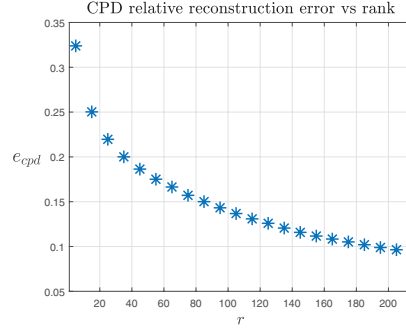[8]Alternating least squares.

[9]CORe CONsistency DIAgnostic.

FIG. 4. *Relative reconstruction error of CPD calculated for different values of rank r for MRI data. As the rank increases, the error becomes smaller.*

5.2.2. Compressed least squares performance. Let $\mathbf{y}_k^{(j)}$ be known in

$$\mathcal{X} \approx \sum_{k=1}^{r} \alpha_k \bigcirc_{j=1}^{d} \mathbf{y}_k^{(j)}$$

for $k \in [r]$ and $j \in [d]$. They can be obtained from a previous iteration in the CPD fitting procedure. Here, they come from the CPD of the data calculated in section 5.2.1. Also, assume these vectors have unit norms. In general, as stated in section 5.2.1, when $\mathbf{y}_k^{(j)}$ are obtained using a CPD algorithm, they do not necessarily have unit norms. Therefore, they are normalized, and the norms are absorbed into the coefficients of CPD. In other words, $\alpha_k = \prod_{j=1}^{d} \|\mathbf{y}_k^{(j)}\|_2$ for $k \in [r]$. If the normalization of the vectors is not performed, $\alpha_k = 1$ for $k \in [r]$. The coefficients of the CPD fit are the solutions to the following least squares problem:

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\beta}} \left\| \mathcal{X} - \sum_{k=1}^{r} \beta_k \bigcirc_{j=1}^{d} \mathbf{y}_k^{(j)} \right\|.$$

As normalization of $\mathbf{y}_k^{(j)}$ was not performed when computing the CPD of the data in these experiments, the true solution will be $\boldsymbol{\alpha} = \mathbf{1}$. An approximate solution for the coefficients can be obtained by solving for

$$\boldsymbol{\alpha}_P = \arg\min_{\boldsymbol{\beta}} \left\| L(\mathcal{X}) - L\left(\sum_{k=1}^{r} \beta_k \bigcirc_{j=1}^{d} \mathbf{y}_k^{(j)} \right) \right\|,$$

where $\boldsymbol{\alpha}_P$ is the vector $\boldsymbol{\alpha}$ estimated for randomly projected data, and $L(\mathcal{X})$ is defined as per (5.1). This is, in fact, simply another way of demonstrating that solving (4.3) yields an approximate solution to (4.2) for a $(d-1)$-mode tensor. Of course, both of these problems can be solved using the vectorized versions of the tensors instead. Indeed, for $\boldsymbol{\alpha}_P$, vectorization should be done after random projection of $\mathcal{X}$ and the rank-1 tensors, i.e.,

$$\boldsymbol{\alpha}_P = \arg\min_{\boldsymbol{\beta}} \|\mathbf{x}_P - \mathbf{B}\boldsymbol{\beta}\|_2 = (\mathbf{B}^*\mathbf{B})^{-1} \mathbf{B}^*\mathbf{x}_P,$$
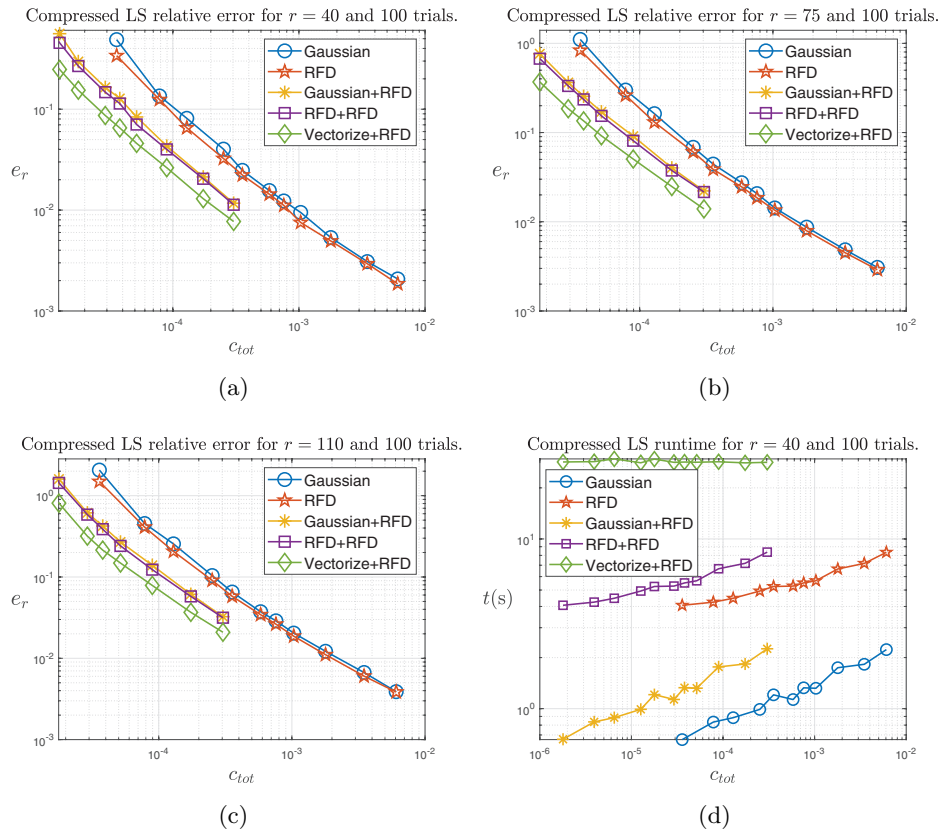
FIG. 5. *Effect of JL embeddings on the relative reconstruction error of least squares (LS) estimation of CPD coefficients. In the 2-stage cases, $c_2 = 0.05$ has been used. (a) $r = 40$. (b) $r = 75$. (c) $r = 110$. (d) Average runtime for $r = 40$. The other runtime plots for $r = 75$ and $r = 110$ are qualitatively identical.*

where $\mathbf{x}_P = \text{vect}\left(L\left(\mathcal{X}\right)\right)$, and $\mathbf{B}$ is a matrix whose k column is $\text{vect}(L(\bigcirc_{j=1}^{d}\mathbf{y}_k^{(j)}))$ for $k \in [r]$. [10, 11] The error measure used to evaluate the approximate solution is defined as

$$e_r = \left|\frac{e_P - e_T}{e_T}\right|,$$

where $e_T = \|\mathcal{X} - \sum_{k=1}^{r}\alpha_k \bigcirc_{j=1}^{d}\mathbf{y}_k^{(j)}\|$ and $e_P = \|\mathcal{X} - \sum_{k=1}^{r}\alpha_{P,k} \bigcirc_{j=1}^{d}\mathbf{y}_k^{(j)}\|$. This, in fact, compares the true CPD reconstruction error and the reconstruction error calculated using the approximate solution for the CPD coefficients $\boldsymbol{\alpha}_P$. The results are shown in Figure 5.

6. Conclusion. We have proposed general modewise Johnson–Lindenstrauss (JL) subspace embeddings that can be both generated much faster and stored more easily than traditional JL embeddings—especially for tensors in very large dimensions.

---

[10] Again, it is clear that in the 2-stage case, $L(\mathcal{X})$ and $L(\bigcirc_{j=1}^{d}\mathbf{y}_k^{(j)})$ are vectors, and therefore the operator vect $(\cdot)$ does not change the result.

[11] The backslash operator was used to actually solve the resulting least squares problems in MATLAB.

We provided a subspace embedding result with improved space complexity bounds for embeddings of rank-r tensors in the setting of unknown basis tensors. This result also has applications in the vector setting, leading to general near-optimal oblivious subspace embedding constructions that require fewer random bits for subspaces spanned by basis vectors having special Kronecker structure. We also provided new fast JL embeddings for arbitrary r-dimensional subspaces using fewer random bits than standard methods. We showcased these results for applications, including compressed least squares and fitting low-rank CP decompositions, while also confirming our results experimentally. There are several interesting future directions, including the analysis of other randomly constructed embeddings, the construction of embeddings designed to maintain other types of structures (such as properties of the core tensor), and their effectiveness in reconstruction and inference tasks.

Appendix A. Proofs of the tensor properties and JL results from section 2. In this section, we give the proofs of Lemmas 1, 2, and 3. The first result lists classical tensor properties that we constantly rely on in this paper.

Proof of Lemma 1. The first property follows from the fact that

$$((\alpha\mathcal{A} + \beta\mathcal{B}) \bigcirc \mathcal{C})_{i_1,\ldots,i_d,i'_1,\ldots,i'_{d'}} = (\alpha\mathcal{A} + \beta\mathcal{B})_{i_1,\ldots,i_d}\, \mathcal{C}_{i'_1,\ldots,i'_{d'}}$$
$$= (\alpha\mathcal{A}_{i_1,\ldots,i_d} + \beta\mathcal{B}_{i_1,\ldots,i_d})\,\mathcal{C}_{i'_1,\ldots,i'_{d'}}.$$

To establish property (ii) we note that

$$\langle \mathcal{A} \bigcirc \mathcal{C}, \mathcal{B} \bigcirc \mathcal{D}\rangle = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \sum_{i'_1=1}^{n'_1} \cdots \sum_{i'_{d'}=1}^{n'_{d'}} \mathcal{A}_{i_1,i_2,\ldots,i_d}\mathcal{C}_{i'_1,\ldots,i'_{d'}}\overline{\mathcal{B}_{i_1,i_2,\ldots,i_d}}\ \overline{\mathcal{D}_{i'_1,\ldots,i'_{d'}}}$$
$$= \left(\sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathcal{A}_{i_1,i_2,\ldots,i_d}\overline{\mathcal{B}_{i_1,i_2,\ldots,i_d}}\right)\left(\sum_{i'_1=1}^{n'_1} \cdots \sum_{i'_{d'}=1}^{n'_{d'}} \mathcal{C}_{i'_1,\ldots,i'_{d'}}\overline{\mathcal{D}_{i'_1,\ldots,i'_{d'}}}\right)$$
$$= \langle \mathcal{A}, \mathcal{B}\rangle \langle \mathcal{C}, \mathcal{D}\rangle.$$

The facts (iii), (iv), and (vi) are easily established using mode-j unfoldings formula (2.5). To establish (iii), we note that

$$((\alpha\mathcal{A} + \beta\mathcal{B}) \times_j \mathbf{U}_j)_{(j)} = \mathbf{U}_j\, (\alpha\mathcal{A} + \beta\mathcal{B})_{(j)} = \mathbf{U}_j\left(\alpha\mathbf{A}_{(j)} + \beta\mathbf{B}_{(j)}\right)$$
$$= \alpha\mathbf{U}_j\mathbf{A}_{(j)} + \beta\mathbf{U}_j\mathbf{B}_{(j)} = \alpha(\mathcal{A} \times_j \mathbf{U}_j)_{(j)} + \beta(\mathcal{B} \times_j \mathbf{U}_j)_{(j)}.$$

Reshaping both sides of the derived equality back into their original tensor forms now completes the proof.[12] The proof of (iv) using unfoldings is nearly identical. To prove (vi) we may again use mode-j unfoldings to see that

$$(\mathcal{A} \times_j \mathbf{U}_j \times_j \mathbf{W})_{(j)} = \mathbf{W}\,(\mathcal{A} \times_j \mathbf{U}_j)_{(j)} = \mathbf{W}\mathbf{U}_j\mathbf{A}_{(j)} = (\mathcal{A} \times_j \mathbf{W}\mathbf{U}_j)_{(j)}.$$

Reshaping these expressions back into their original tensor forms again completes the proof. To prove (v), it is perhaps easiest to appeal directly to the componentwise definition of the mode-j product given in (2.4). Suppose that $\ell > j$ (the case $\ell < j$ is

---

[12]Here we are implicitly using the fact that mode-j unfolding provides a vector space isomorphism between $\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ and $\mathbb{C}^{n_j \times \prod_{\ell \in [d]\setminus\{j\}} n_\ell}$ for all $j \in [d]$.

nearly identical). Set $\mathbf{U} := \mathbf{U}_j$ and $\mathbf{V} := \mathbf{V}_\ell$ to simplify the subscript notation. We have for all $k \in [m_j]$, $l \in [m_\ell]$, and $i_q \in [n_q]$ with $q \notin \{j, \ell\}$ that

$$((\mathcal{A} \times_j \mathbf{U}) \times_\ell \mathbf{V})_{i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_{\ell-1},l,i_{\ell+1},\ldots,i_d}$$

$$= \sum_{i_\ell=1}^{n_\ell} (\mathcal{A} \times_j \mathbf{U})_{i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_\ell,\ldots,i_d} \mathbf{V}_{l,i_\ell}$$

$$= \sum_{i_\ell=1}^{n_\ell} \left(\sum_{i_j=1}^{n_j} \mathcal{A}_{i_1,\ldots,i_j,\ldots,i_\ell,\ldots,i_d} \mathbf{U}_{k,i_j} \right) \mathbf{V}_{l,i_\ell}$$

$$= \sum_{i_j=1}^{n_j} \left(\sum_{i_\ell=1}^{n_\ell} \mathcal{A}_{i_1,\ldots,i_j,\ldots,i_\ell,\ldots,i_d} \mathbf{V}_{l,i_\ell} \right) \mathbf{U}_{k,i_j}$$

$$= \sum_{i_j=1}^{n_j} (\mathcal{A} \times_\ell \mathbf{V})_{i_1,\ldots,i_j,\ldots,i_{\ell-1},l,i_{\ell+1},\ldots,i_d} \mathbf{U}_{k,i_j}$$

$$= ((\mathcal{A} \times_\ell \mathbf{U}) \times_j \mathbf{U})_{i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_{\ell-1},l,i_{\ell+1},\ldots,i_d}. \qquad \square$$

Our second proof of this appendix shows that JL embeddings can also preserve the inner products between all elements of a given finite set.

Proof of Lemma 2. The result for vectors is a well-known consequence of the polarization identity for inner products. We have that

$$|\langle \mathbf{Ax},\ \mathbf{Ay} \rangle - \langle \mathbf{x},\ \mathbf{y} \rangle| = \left| \frac{1}{4} \sum_{\ell=0}^{3} \mathrm{i}^\ell \left(\|\mathbf{Ax} + \mathrm{i}^\ell \mathbf{Ay}\|_2^2 - \|\mathbf{x} + \mathrm{i}^\ell \mathbf{y}\|_2^2 \right) \right|$$

$$= \left| \frac{1}{4} \sum_{\ell=0}^{3} \mathrm{i}^\ell \varepsilon_\ell \|\mathbf{x} + \mathrm{i}^\ell \mathbf{y}\|_2^2 \right|$$

$$\leq \frac{1}{4} \sum_{\ell=0}^{3} \varepsilon (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2 = \varepsilon (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2$$

$$= \varepsilon \left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + 2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \right)$$

$$\leq 2\varepsilon \left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 \right) \leq 4\varepsilon \cdot \max \left\{ \|\mathbf{x}\|_2^2, \|\mathbf{y}\|_2^2 \right\},$$

where the second-to-last inequality follows from Young's inequality for products. The proof of the tensor counterpart is essentially identical, with $L(\mathcal{X})$ replacing $\mathbf{Ax}$, and making use of the linearity of L. $\qquad \square$

The next proof is a version of classical covering estimate in high-dimensional spaces.

Proof of Lemma 3. The cardinality bound on $\mathcal{C}$ can be obtained from the covering results in Appendix C of [23]. It is enough to establish (2.9) for an arbitrary $\mathbf{x} \in \mathcal{S}_{\ell^2}$ due to the linearity of $\mathbf{A}$ and $\mathcal{L}$. Let $\Delta := \|\mathbf{A}\|_{2\to2} \geq 0$, and choose an element $\mathbf{y} \in \mathcal{C}$ with $\|\mathbf{x} - \mathbf{y}\| \leq \varepsilon/16$. We have that

$$\|\mathbf{Ax}\|_2 - \|\mathbf{x}\|_2 \leq \|\mathbf{Ay}\|_2 + \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2 - 1 \leq \sqrt{1 + \varepsilon/2} - 1 + \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2$$

$$\leq (1 + \varepsilon/4) - 1 + \Delta\varepsilon/16 = (\varepsilon/4)(1 + \Delta/4)$$

holds for all $\mathbf{x} \in \mathcal{S}_{\ell^2}$. This, in turn, means that the upper bound above will hold for a vector $\mathbf{x}$ realizing $\|\mathbf{Ax}\| = \|\mathbf{A}\|_{2\to2}$ so that $\Delta - 1 \leq (\varepsilon/4)(1 + \Delta/4)$ must also hold.

As a consequence, $\Delta \leq 1 + \varepsilon/4 + \Delta\varepsilon/16 \implies \Delta \leq \frac{1+\varepsilon/4}{1-\varepsilon/16} \leq 1 + \varepsilon/3$. The upper bound now follows.

To establish the lower bound we define $\delta := \inf_{\mathbf{z} \in \mathcal{S}_{\ell^2}} \|\mathbf{A}\mathbf{z}\| \geq 0$ and note that this quantity will also be realized by some element of the compact set $\mathcal{S}_{\ell^2}$. As above we consider this minimizing vector $\mathbf{x} \in \mathcal{S}_{\ell^2}$ and choose an element $\mathbf{y} \in \mathcal{C}$ with $\|\mathbf{x} - \mathbf{y}\| \leq \varepsilon/16$ in order to see that

$$\begin{aligned} \delta - 1 = \|\mathbf{A}\mathbf{x}\|_2 - \|\mathbf{x}\|_2 &\geq \|\mathbf{A}\mathbf{y}\|_2 - \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2 - 1 \geq \sqrt{1 - \varepsilon/2} - 1 - \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2 \\ &\geq (1 - \varepsilon/3) - 1 - \Delta\varepsilon/16 \geq -(\varepsilon/3 + \varepsilon/16\,(1 + \varepsilon/3)) \\ &\geq -(\varepsilon/3 + \varepsilon/16 + \varepsilon/48) = -5\varepsilon/12. \end{aligned}$$

As a consequence, $\delta \geq 1 - 5\varepsilon/12$. The lower bound now follows. $\square$

Appendix B. Proofs of the intermediate results from sections 3.1 and 4.1.

In this section, we give the proofs of all auxiliary results for the proof of Theorem 3. All the statements are listed in section 3.1.

Proof of Lemma 5. Using Lemma 1, the linearity of tensor matricization, and (2.7) we can see that the mode-j unfolding of $\mathcal{Y}'$ satisfies

$$\begin{aligned} \mathbf{Y'}_{(j)} = \mathbf{B}\mathbf{Y}_{(j)} &= \mathbf{B}\sum_{k=1}^{r} \alpha_k \left(\bigcirc_{\ell=1}^{d} \mathbf{y}_k^{(\ell)}\right)_{(j)} = \sum_{k=1}^{r} \alpha_k \mathbf{B}\mathbf{y}_k^{(j)} \left(\otimes_{\ell\neq j} \mathbf{y}_k^{(\ell)}\right)^{\top} \\ &= \sum_{k=1}^{r} \left(\alpha_k \left\|\mathbf{B}\mathbf{y}_k^{(j)}\right\|_2\right) \frac{\mathbf{B}\mathbf{y}_k^{(j)}}{\left\|\mathbf{B}\mathbf{y}_k^{(j)}\right\|_2} \left(\otimes_{\ell\neq j} \mathbf{y}_k^{(\ell)}\right)^{\top}. \end{aligned}$$

Refolding $\mathbf{Y'}_{(j)}$ back into a d-mode tensor then gives us our first equality. The next two equalities now follow directly from the definitions of modewise coherence. $\square$

Proof of Lemma 6. Using Lemma 1, the linearity of tensor matricization, and (2.7), once again, we can see that

$$\begin{aligned} \|\mathcal{Y} \times_j \mathbf{B}\|^2 &= \left\|\sum_{k=1}^{r} \alpha_k \left(\bigcirc_{\ell=1}^{d} \mathbf{y}_k^{(\ell)} \times_j \mathbf{B}\right)\right\|^2 = \left\|\sum_{k=1}^{r} \alpha_k \mathbf{B}\mathbf{y}_k^{(j)} \left(\otimes_{\ell\neq j} \mathbf{y}_k^{(\ell)}\right)^{\top}\right\|_{\mathrm{F}}^2 \\ &= \sum_{k,h=1}^{r} \left\langle \alpha_k \mathbf{B}\mathbf{y}_k^{(j)} \left(\otimes_{\ell\neq j} \mathbf{y}_k^{(\ell)}\right)^{\top}, \alpha_h \mathbf{B}\mathbf{y}_h^{(j)} \left(\otimes_{\ell\neq j} \mathbf{y}_h^{(\ell)}\right)^{\top}\right\rangle_{\mathrm{F}}, \end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ and $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ denote the Frobenius matrix norm and inner product, respectively. Computing the Frobenius inner products above columnwise by expressing each $\mathbf{B}\mathbf{y}_k^{(j)}(\otimes_{\ell\neq j}\mathbf{y}_k^{(\ell)})^{\top}$ as a sum of its individual columns (each represented as a matrix with only one nonzero column) we can further see that

$$\|\mathcal{Y} \times_j \mathbf{B}\|^2 = \sum_{k,h=1}^{r} \sum_{a=1}^{\prod_{\ell\neq j} n_\ell} \alpha_k \left(\otimes_{\ell\neq j} \mathbf{y}_k^{(\ell)}\right)_a \overline{\alpha_h \left(\otimes_{\ell\neq j} \mathbf{y}_h^{(\ell)}\right)_a} \left\langle \mathbf{B}\mathbf{y}_k^{(j)}, \mathbf{B}\mathbf{y}_h^{(j)}\right\rangle$$

as we wished to show. $\square$

Proof of Proposition 1. We prove each property in order below.

Proof of (†). By Lemma 5 we have for all $k \in [r]$ that

$$\left| \alpha_k' - \alpha_k \right| \;=\; \left| \alpha_k \left\| \mathbf{A}\mathbf{y}_k^{(j)} \right\|_2 - \alpha_k \right| \;=\; \left| \left\| \mathbf{A}\mathbf{y}_k^{(j)} \right\|_2 - 1 \right| |\alpha_k| \;\leq\; \varepsilon |\alpha_k|/4$$

as we wished to prove.

Proof of (††). Appealing to Lemma 5 and the definition of j-mode coherence, we have that

$$\mu_{\mathcal{Y}',j} = \max_{\substack{k,h \in [r] \\ k \neq h}} \frac{\left| \left\langle \mathbf{A}\mathbf{y}_k^{(j)}, \mathbf{A}\mathbf{y}_h^{(j)} \right\rangle \right|}{\left\| \mathbf{A}\mathbf{y}_k^{(j)} \right\|_2 \left\| \mathbf{A}\mathbf{y}_h^{(j)} \right\|_2} \;\leq\; \max_{\substack{k,h \in [r] \\ k \neq h}} \frac{\left| \left\langle \mathbf{y}_k^{(j)}, \mathbf{y}_h^{(j)} \right\rangle \right| + \varepsilon}{1 - \frac{\varepsilon}{4}} = \frac{\mu_{\mathcal{Y},j} + \varepsilon}{1 - \frac{\varepsilon}{4}},$$

where the inequality follows from Lemma 2 combined with $\mathbf{A}$ being an $(\varepsilon/4)$-JL embedding.

Proof of († † †). Applying Lemma 6 with $\mathbf{B} = \mathbf{A}$ and $\mathbf{B} = \mathbf{I}$, respectively, we can see that

(B.1)

$$\begin{aligned} &\|\mathcal{Y}'\|^2 - \|\mathcal{Y}\|^2 \\ &= \sum_{k,h=1}^{r} \sum_{a=1}^{\prod_{\ell \neq j} n_\ell} \alpha_k \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)_a \overline{\alpha_h \left(\otimes_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right)_a} \left(\left\langle \mathbf{A}\mathbf{y}_k^{(j)}, \mathbf{A}\mathbf{y}_h^{(j)} \right\rangle - \left\langle \mathbf{y}_k^{(j)}, \mathbf{y}_h^{(j)} \right\rangle \right). \end{aligned}$$

Applying Lemma 2 to each inner product in (16), we can now see that

$$\left\langle \mathbf{A}\mathbf{y}_k^{(j)}, \mathbf{A}\mathbf{y}_h^{(j)} \right\rangle \;=\; \left\langle \mathbf{y}_k^{(j)}, \mathbf{y}_h^{(j)} \right\rangle + \varepsilon_{k,h}$$

for some $\varepsilon_{k,h} \in \mathbb{C}$ with $|\varepsilon_{k,h}| \leq \varepsilon$. As a result we have that

$$\begin{aligned} \left| \|\mathcal{Y} \times_j \mathbf{A}\|^2 - \|\mathcal{Y}\|^2 \right| &= \left| \sum_{k,h=1}^{r} \sum_{a=1}^{\prod_{\ell \neq j} n_\ell} \alpha_k \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)_a \overline{\alpha_h \left(\otimes_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right)_a} \varepsilon_{k,h} \right| \\ &= \left| \sum_{k,h=1}^{r} \alpha_k \overline{\alpha_h} \varepsilon_{k,h} \sum_{a=1}^{\prod_{\ell \neq j} n_\ell} \left(\otimes_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right)_a \overline{\left(\otimes_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right)_a} \right| \\ &= \left| \sum_{k,h=1}^{r} \alpha_k \overline{\alpha_h} \varepsilon_{k,h} \left\langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right\rangle \right| \\ &\leq \left| \sum_{k=1}^{r} |\alpha_k|^2 \varepsilon_{k,k} \left\| \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right\|^2 \right| \\ &\quad + \left| \sum_{k \neq h} \alpha_k \overline{\alpha_h} \varepsilon_{k,h} \left\langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right\rangle \right|. \end{aligned}$$

Noting that $\left\| \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)} \right\|^2 = 1$ by Lemma 1 since $\|\mathbf{y}_k^{(\ell)}\|_2 = 1$ for all $\ell \in [d]$ and

$k \in [r]$, we now have that

$$\left| \|\mathcal{Y} \times_j \mathbf{A}\|^2 - \|\mathcal{Y}\|^2 \right| \;\leq\; \varepsilon \left| \sum_{k=1}^{r} |\alpha_k|^2 \right| + \left| \sum_{k \neq h} \alpha_k \overline{\alpha_h} \varepsilon_{k,h} \left\langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right\rangle \right|$$

$$= \; \varepsilon \|\boldsymbol{\alpha}\|_2^2 + \left| \left\langle \mathbf{E}^\top \boldsymbol{\alpha}, \boldsymbol{\alpha} \right\rangle \right|,$$

where $\mathbf{E} \in \mathbb{C}^{r \times r}$ is zero on its diagonal, and $E_{k,h} = \varepsilon_{k,h} \langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \rangle$ for $k \neq h$. As a result, $\left| \|\mathcal{Y} \times_j \mathbf{A}\|^2 - \|\mathcal{Y}\|^2 \right| \leq \left(\varepsilon + \left\| \mathbf{E}^\top \right\|_{2\to2} \right) \|\boldsymbol{\alpha}\|_2^2$, where the operator norm $\left\| \mathbf{E}^\top \right\|_{2\to2}$ satisfies

$$\left\| \mathbf{E}^\top \right\|_{2\to2} \;\leq\; \|\mathbf{E}\|_F \leq \sqrt{\sum_{k \neq h} \left| \left\langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right\rangle \right|^2 \varepsilon^2}$$

$$= \varepsilon \; \sqrt{\sum_{k \neq h} \left| \left\langle \bigcirc_{\ell \neq j} \mathbf{y}_k^{(\ell)}, \bigcirc_{\ell \neq j} \mathbf{y}_h^{(\ell)} \right\rangle \right|^2}.$$

Finally, Lemma 1 and the definition of $\mu_{\mathcal{Y}}$ imply that

$$\|\mathbf{E}\|_{2\to2} \;\leq\; \varepsilon \sqrt{r(r-1)} \prod_{\ell \neq j} \mu_{\mathcal{Y},\ell} \;\leq\; \varepsilon r \mu_{\mathcal{Y}}^{d-1}.$$

Thus, we obtain the desired bound,

$$\left| \|\mathcal{Y} \times_j \mathbf{A}\|^2 - \|\mathcal{Y}\|^2 \right| \;\leq\; \varepsilon \left(1 + \sqrt{r(r-1)} \prod_{\ell \neq j} \mu_{\mathcal{Y},\ell} \right) \|\boldsymbol{\alpha}\|_2^2 \leq \varepsilon \left(1 + r \mu_{\mathcal{Y}}^{d-1} \right) \|\boldsymbol{\alpha}\|_2^2. \quad \square$$

Proof of Lemma 7. Utilizing Lemma 1 and the standard form of $\mathcal{Y}$, we can see that

$$\left| \|\mathcal{Y}\|^2 - \|\boldsymbol{\alpha}\|_2^2 \right| \;=\; \left| \sum_{k,h=1}^{r} \alpha_k \overline{\alpha_h} \left\langle \bigcirc_{\ell=1}^{d} \mathbf{y}_k^{(\ell)}, \; \bigcirc_{\ell=1}^{d} \mathbf{y}_h^{(\ell)} \right\rangle - \sum_{k=1}^{r} |\alpha_k|^2 \right|$$

$$= \; \left| \sum_{k \neq h}^{r} \alpha_k \overline{\alpha_h} \prod_{\ell=1}^{d} \left\langle \mathbf{y}_k^{(\ell)}, \; \mathbf{y}_h^{(\ell)} \right\rangle \right| \;\leq\; \mu_{\mathcal{Y}}' \sum_{k \neq h}^{r} |\alpha_k \overline{\alpha_h}|$$

$$= \; \mu_{\mathcal{Y}}' \left(\left(\sum_{k=1}^{r} |\alpha_k| \right)^2 - \sum_{k=1}^{r} |\alpha_k|^2 \right) \;\leq\; \mu_{\mathcal{Y}}' \left(\left(\sqrt{r} \|\boldsymbol{\alpha}\|_2 \right)^2 - \|\boldsymbol{\alpha}\|_2^2 \right),$$

where the last inequality follows from the Cauchy–Schwarz inequality. As a result we have that

$$\left| \|\mathcal{Y}\|^2 - \|\boldsymbol{\alpha}\|_2^2 \right| \;\leq\; \mu_{\mathcal{Y}}'(r-1)\|\boldsymbol{\alpha}\|_2^2,$$

which in turn implies that

$$\|\mathcal{Y}\|^2 \geq \left(1 - (r-1)\mu_{\mathcal{Y}}' \right) \|\boldsymbol{\alpha}\|_2^2. \quad \square$$

The following simple fact will be used repeatedly in the proof of Proposition 2.

Remark 7. Let $c, d \in \mathbb{R}^+$. Then, $\mathbb{e}^c \geq (1 + \frac{c}{d})^d$.

Proof of Proposition 2. Let $\mathcal{Y}^{(0)} := \mathcal{Y}$, and for each $j \in [d]$ define the tensor

$$\mathcal{Y}^{(j)} := \mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_j \mathbf{A}_j \ = \ \sum_{k=1}^{r} \alpha_{j,k} \bigcirc_{\ell=1}^{d} \mathbf{y}_{j,k}^{(\ell)}$$

expressed in standard form via j applications of Lemma 5. Note that parts (†) and (††) of Proposition 1 imply that both of the following hold for all $j \in [d]$:

(i) $|\alpha_{j,k} - \alpha_{j-1,k}| \leq \varepsilon|\alpha_{j-1,k}|/4d$ so that $|\alpha_{j,k}| \leq (1 + \varepsilon/4d)|\alpha_{j-1,k}|$ holds for all $k \in [r]$, and

(ii) $\mu_{\mathcal{Y}^{(j)},j} \leq (\mu_{\mathcal{Y}^{(j-1)},j} + \varepsilon/d)/(1 - \varepsilon/4d)$, and $\mu_{\mathcal{Y}^{(j)},\ell} = \mu_{\mathcal{Y}^{(j-1)},\ell}$ for all $\ell \in [d] \setminus \{j\}$.

Using these facts, it is not too difficult to inductively establish that both

(B.2) $$|\alpha_{j,k}| \leq (1 + \varepsilon/4d)^j |\alpha_k|$$

and

(B.3) $$\prod_{\ell \neq j} \mu_{\mathcal{Y}^{(j-1)},\ell} \ \leq \ \left(\prod_{\ell < j} \frac{\mu_{\mathcal{Y},\ell} + \varepsilon/d}{1 - \varepsilon/4d} \right) \prod_{\ell > j} \mu_{\mathcal{Y},\ell} \ \leq \ \left(\frac{\mu_{\mathcal{Y}} + \varepsilon/d}{1 - \varepsilon/4d} \right)^{j-1} \mu_{\mathcal{Y}}^{d-j}$$

also hold for all $k \in [r]$ and $j \in [d]$. Note that in (B.3) we will let $\mu_{\mathcal{Y}}^0 = 1$ even if $\mu_{\mathcal{Y}} = 0$ since this still yields the correct bound in the case where $j = d$ and $\mu_{\mathcal{Y}} = 0$.

Proceeding with the desired error bound, we can now see that

$$\left| \|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2 \right| = \left| \sum_{j=0}^{d-1} \left\| \mathcal{Y}^{(j)} \right\|^2 - \left\| \mathcal{Y}^{(j+1)} \right\|^2 \right|$$

$$\leq \frac{\varepsilon}{d} \sum_{j=0}^{d-1} \left(1 + \sqrt{r(r-1)} \prod_{\ell \neq j+1} \mu_{\mathcal{Y}^{(j)},\ell} \right) \|\boldsymbol{\alpha}_j\|_2^2$$

$$\leq \frac{\varepsilon}{d} \sum_{j=0}^{d-1} \left(1 + \sqrt{r(r-1)} \left(\frac{\mu_{\mathcal{Y}} + \varepsilon/d}{1 - \varepsilon/4d} \right)^j \mu_{\mathcal{Y}}^{d-1-j} \right) (1 + \varepsilon/4d)^{2j} \|\boldsymbol{\alpha}\|_2^2$$

$$\leq \frac{\varepsilon}{d} \sum_{j=0}^{d-1} \left(1 + \sqrt{r(r-1)} \left(\frac{\mu_{\mathcal{Y}} + \varepsilon/d}{1 - \varepsilon/4d} \right)^j \mu_{\mathcal{Y}}^{d-1-j} \right) (1 + 9\varepsilon/16d)^j \|\boldsymbol{\alpha}\|_2^2,$$

where we have used part († † †) of Proposition 1, (B.2), and (B.3). Considering each term in the upper bound above separately, we have that

$$\left| \|\mathcal{Y}\|^2 - \|\mathcal{Y} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d\|^2 \right| \leq \frac{\varepsilon}{d} \|\boldsymbol{\alpha}\|_2^2 \left(T_1 + \sqrt{r(r-1)} T_2 \right),$$

where

$$T_1 := \sum_{j=0}^{d-1} (1 + 9\varepsilon/16d)^j \ = \ \frac{(1 + 9\varepsilon/16d)^d - 1}{9\varepsilon/16d} \ \leq \ \mathrm{e}d,$$

using Remark 7 and $9\varepsilon/16 < 1$, and where

$$T_2 := \sum_{j=0}^{d-1} \left(\frac{\mu_{\mathcal{Y}} + \varepsilon/d}{1 - \varepsilon/4d} \right)^j \mu_{\mathcal{Y}}^{d-1-j} (1 + 9\varepsilon/16d)^j \ \leq \ \sum_{j=0}^{d-1} (\mu_{\mathcal{Y}} + \varepsilon/d)^j \mu_{\mathcal{Y}}^{d-1-j} (1 + \varepsilon/d)^j$$

for $\varepsilon \leq 3/4$.

Continuing to bound the second term we will consider three cases. First, if $\mu_{\mathcal{Y}} = 0$, then
$$T_2 \ \leq \ (\varepsilon/d)^{d-1} (1 + \varepsilon/d)^{d-1} \ \leq \ \mathrm{e}\,(\varepsilon/d)^{d-1}$$
using Remark 7 and $\varepsilon < 1$. Second, if $0 < \mu_{\mathcal{Y}} \leq \varepsilon$, then
$$\begin{aligned} T_2 \ &\leq \ \sum_{j=0}^{d-1} (\varepsilon + \varepsilon/d)^j \, \varepsilon^{d-1-j} (1 + \varepsilon/d)^j \ = \ \varepsilon^{d-1} \sum_{j=0}^{d-1} (1 + 1/d)^j \, (1 + \varepsilon/d)^j \\ &\leq \ \varepsilon^{d-1} d \, (1 + 1/d)^d \, (1 + \varepsilon/d)^d \ \leq \ d\mathrm{e}^2 \varepsilon^{d-1}, \end{aligned}$$
using Remark 7 and $\varepsilon < 1$ once more. If, however, $\mu_{\mathcal{Y}} > \varepsilon$, then we can see that
$$\begin{aligned} T_2 \ &\leq \ \mu_{\mathcal{Y}}^{d-1} \sum_{j=0}^{d-1} (1 + \varepsilon/\mu_{\mathcal{Y}} d)^j \, (1 + \varepsilon/d)^j \ \leq \ \mu_{\mathcal{Y}}^{d-1} \sum_{j=0}^{d-1} (1 + 1/d)^j \, (1 + \varepsilon/d)^j \\ &\leq \ \mu_{\mathcal{Y}}^{d-1} \cdot d \, (1 + 1/d)^d \, (1 + \varepsilon/d)^d \ \leq \ \mu_{\mathcal{Y}}^{d-1} \, d \, \mathrm{e}^{1+\varepsilon} \ \leq \ d\mathrm{e}^2 \mu_{\mathcal{Y}}^{d-1}, \end{aligned}$$
where we have again utilized Remark 7. The desired result now follows.   □

Proof of Lemma 9. Fix $t \in [p]$ and let $\mathcal{X}^{(0)} := \mathcal{Z}^{(t)}$, $\mathcal{X}^{(j)} := \mathcal{Z}^{(j,t)}$ for all $j \in [d-1]$, and let $\mathcal{X}^{(d)} := \mathcal{Z}^{(d-1,t)} \times_d \mathbf{A}_d = \mathcal{Z}^{(t)} \times_1 \mathbf{A}_1 \cdots \times_d \mathbf{A}_d$. Choose any $j \in [d]$, and let $\mathbf{x}_{j,h} \in \mathbb{C}^{n_j}$ denote the hth column of the mode-j unfolding of $\mathcal{X}^{(j-1)}$, denoted by $\mathbf{X}^{(j-1)}_{(j)}$. It is easy to see that each $\mathbf{x}_{j,h}$ is a mode-j fiber of $\mathcal{X}^{(j-1)} = \mathcal{Z}^{(j-1,t)}$ for each $1 \leq h \leq N'_j := (\prod_{\ell=1}^{j-1} m_\ell)(\prod_{\ell=j+1}^{d} n_\ell)$. Thus, we can see that

$$\begin{aligned} \left| \left\| \mathcal{X}^{(j-1)} \right\|^2 - \left\| \mathcal{X}^{(j)} \right\|^2 \right| &= \left| \left\| \mathcal{X}^{(j-1)} \right\|^2 - \left\| \mathcal{X}^{(j-1)} \times_j \mathbf{A}_j \right\|^2 \right| \\ &= \left| \left\| \mathbf{X}^{(j-1)}_{(j)} \right\|_{\mathrm{F}}^2 - \left\| \mathbf{A}_j \mathbf{X}^{(j-1)}_{(j)} \right\|_{\mathrm{F}}^2 \right| \\ &= \left| \sum_{h=1}^{N'_j} \| \mathbf{x}_{j,h} \|_2^2 - \| \mathbf{A}_j \mathbf{x}_{j,h} \|_2^2 \right| \ \leq \ \sum_{h=1}^{N'_j} \left| \| \mathbf{x}_{j,h} \|_2^2 - \| \mathbf{A}_j \mathbf{x}_{j,h} \|_2^2 \right| \\ &\leq \ \frac{\varepsilon}{\mathrm{e}d} \sum_{h=1}^{N'_j} \| \mathbf{x}_{j,h} \|_2^2 = \frac{\varepsilon}{\mathrm{e}d} \left\| \mathbf{X}^{(j-1)}_{(j)} \right\|_{\mathrm{F}}^2 = \frac{\varepsilon}{\mathrm{e}d} \left\| \mathcal{X}^{(j-1)} \right\|^2. \end{aligned}$$

A short induction argument now reveals that $\left\| \mathcal{X}^{(j)} \right\|^2 \ \leq \ \left(1 + \frac{\varepsilon}{\mathrm{e}d} \right)^j \left\| \mathcal{X}^{(0)} \right\|^2$ holds for all $j \in [d]$. As a result we can now see that

$$\begin{aligned} \left| \left\| \mathcal{X}^{(0)} \right\|^2 - \left\| \mathcal{X}^{(d)} \right\|^2 \right| &= \left| \sum_{j=1}^{d} \left\| \mathcal{X}^{(j-1)} \right\|^2 - \left\| \mathcal{X}^{(j)} \right\|^2 \right| \\ &\leq \sum_{j=1}^{d} \left| \left\| \mathcal{X}^{(j-1)} \right\|^2 - \left\| \mathcal{X}^{(j)} \right\|^2 \right| \ \leq \ \frac{\varepsilon}{\mathrm{e}d} \sum_{j=1}^{d} \left\| \mathcal{X}^{(j-1)} \right\|^2 \\ &\leq \ \frac{\varepsilon}{\mathrm{e}d} \sum_{j=1}^{d} \left(1 + \frac{\varepsilon}{\mathrm{e}d} \right)^{j-1} \left\| \mathcal{X}^{(0)} \right\|^2 \ \leq \ \frac{\varepsilon}{\mathrm{e}} \left(1 + \frac{\varepsilon}{\mathrm{e}d} \right)^d \left\| \mathcal{X}^{(0)} \right\|^2 \end{aligned}$$

holds. The desired result now follows from Remark 7.   □

Acknowledgments. Mark would like to thank E.I. and D. M. for greatly accentuating his UCLA visit by squatting at his Airbnb Oct. 15–19, 2019, as well as committing a written act of dogeza to his near-optimal wife for agreeing to his being over 2000 miles away during E.'s witching months. Mark also sends many thanks to E. S. for helping out with the baby in his place during his absence.

REFERENCES

[1] *Alzheimer's Disease Neuroimaging Initiative*, http://adni.loni.usc.edu/.

[2] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, J. Comput. System Sci., 66 (2003), pp. 671–687.

[3] T. D. Ahle, M. Kapralov, J. B. T. Knudsen, R. Pagh, A. Velingker, D. P. Woodruff, and A. Zandieh, *Oblivious sketching of high-degree polynomial kernels*, in Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2020, pp. 141–160, https://doi.org/10.1137/1.9781611975994.9.

[4] A. Ahmed and J. Romberg, *Compressive multiplexing of correlated signals*, IEEE Trans. Inform. Theory, 61 (2014), pp. 479–498.

[5] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, *Tensor decompositions for learning latent variable models*, J. Mach. Learn. Res., 15 (2014), pp. 2773–2832.

[6] H. Avron, H. Nguyen, and D. Woodruff, *Subspace embeddings for the polynomial kernel*, in Advances in Neural Information Processing Systems, Vol. 27, Curran Associates, Inc., 2014, pp. 2258–2266.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx., 28 (2008), pp. 253–263.

[8] R. Basri and D. W. Jacobs, *Lambertian reflectance and linear subspaces*, IEEE Trans. Pattern Anal. Mach. Intell., 25 (2003), pp. 218–233.

[9] C. Battaglino, G. Ballard, and T. G. Kolda, *A practical randomized CP tensor decomposition*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 876–901, https://doi.org/10.1137/17M1112303.

[10] M. H. Beck, A. Jäckle, G. A. Worth, and H.-D. Meyer, *The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propagating wavepackets*, Phys. Rep., 324 (2000), pp. 1–105.

[11] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do, *Efficient tensor completion for color image and video recovery: Low-rank tensor train*, IEEE Trans. Image Process., 26 (2017), pp. 2466–2479.

[12] S. Bittens, R. Zhang, and M. A. Iwen, *A deterministic sparse FFT for functions with structured Fourier sparsity*, Adv. Comput. Math., 45 (2019), pp. 519–561.

[13] R. Bro and H. A. Kiers, *A new efficient method for determining the number of components in PARAFAC models*, J. Chemometrics, 17 (2003), pp. 274–286.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM, 58 (2011), 11.

[15] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[16] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[17] J. D. Carroll and J.-J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[18] M. Charikar, K. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, Theoret. Comput. Sci., 312 (2004), pp. 3–15.

[19] G. Cormode and S. Muthukrishnan, *What's hot and what's not: Tracking most frequent items dynamically*, ACM Trans. Database Syst., 30 (2005), pp. 249–278.

[20] A. Dasgupta, R. Kumar, and T. Sarlós, *A sparse Johnson–Lindenstrauss transform*, in Proceedings of the Forty-Second ACM Symposium on Theory of Computing, ACM, 2010, pp. 341–350.

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

[22] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[23] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, 2013.

[24] A. C. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, *Recent developments in the sparse Fourier transform: A compressed Fourier transform for big data*, IEEE Signal Process. Mag., 31 (2014), pp. 91–100.

[25] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, *Group testing and sparse signal recovery*, in Proceedings of the 2008 42nd Asilomar Conference on Signals, Systems and Computers, IEEE, 2008, pp. 1059–1063.

[26] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing*, Phys. Rev. Lett., 105 (2010), 150401.

[27] R. A. Harshman, *Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multimodal Factor Analysis*, Tech. report, University of California at Los Angeles, 1970.

[28] M. A. Iwen and B. W. Ong, *A distributed and incremental SVD algorithm for agglomerative data analysis on large networks*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1699–1718, https://doi.org/10.1137/16M1058467.

[29] M. A. Iwen, *Combinatorial sublinear-time Fourier algorithms*, Found. Comput. Math., 10 (2010), pp. 303–338.

[30] M. A. Iwen, *Improved approximation guarantees for sublinear-time Fourier algorithms*, Appl. Comput. Harmon. Anal., 34 (2013), pp. 57–82.

[31] M. A. Iwen, *Compressed sensing with sparse binary matrices: Instance optimal error guarantees in near-optimal time*, J. Complex., 30 (2014), pp. 1–15.

[32] R. Jin, T. G. Kolda, and R. Ward, *Faster Johnson–Lindenstrauss Transforms via Kronecker Products*, preprint, https://arxiv.org/abs/1909.04801, 2019.

[33] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Proceedings of the Conference in Modern Analysis and Probability (New Haven, Conn., 1982), Contemp. Math. 26, Amer. Math. Soc., 1984, pp. 189–206.

[34] D. M. Kane and J. Nelson, *Sparser Johnson-Lindenstrauss transforms*, J. ACM, 61 (2014), 4.

[35] T. G. Kolda, *Orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 243–255, https://doi.org/10.1137/S0895479800368354.

[36] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500, https://doi.org/10.1137/07070111X.

[37] F. Krahmer and R. Ward, *New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal., 43 (2011), pp. 1269–1281, https://doi.org/10.1137/100810447.

[38] K. G. Larsen and J. Nelson, *Optimality of the Johnson-Lindenstrauss lemma*, in Proceedings of the 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2017, pp. 633–638.

[39] X. Li, J. Haupt, and D. Woodruff, *Near optimal sketching of low-rank tensor regression*, in Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 3466–3476.

[40] J. Liu, P. Musialski, P. Wonka, and J. Ye, *Tensor completion for estimating missing values in visual data*, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2012), pp. 208–220.

[41] C. Lubich, *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*, Zur. Lect. Adv. Math., European Mathematical Society, Zürich, 2008.

[42] O. A. Malik and S. Becker, *Low-rank Tucker decomposition of large tensors using tensorsketch*, in Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018, pp. 10096–10106.

[43] O. A. Malik and S. Becker, *Guarantees for the Kronecker Fast Johnson-Lindenstrauss Transform Using a Coherence and Sampling Argument*, preprint, https://arxiv.org/abs/1911.08424, 2019.

[44] S. Merhi, R. Zhang, M. A. Iwen, and A. Christlieb, *A new class of fully discrete sparse Fourier transforms: Faster stable implementations with guarantees*, J. Fourier Anal. Appl., 25 (2019), pp. 751–784.

[45] R. Pagh, *Compressed matrix multiplication*, ACM Trans. Comput. Theory, 5 (2013), pp. 1–17.

[46] N. Pham and R. Pagh, *Fast and scalable polynomial kernels via explicit feature maps*, in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 239–247.

[47] B. T. Rakhshan and G. Rabusseau, *Tensorized Random Projections*, preprint https://arxiv.org/abs/2003.05101, 2020.

[48] H. Rauhut, R. Schneider, and Ž. Stojanac, *Low rank tensor recovery via iterative hard*

thresholding, Linear Algebra Appl., 523 (2017), pp. 220–262.

[49] B. RECHT, M. FAZEL, AND P. A. PARRILO, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501, https://doi.org/10.1137/070697835.

[50] B. ROMERA-PAREDES, H. AUNG, N. BIANCHI-BERTHOUZE, AND M. PONTIL, *Multilinear multi-task learning*, in Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, PMLR 28, 2013, pp. 1444–1452.

[51] B. SEGAL AND M. IWEN, *Improved sparse Fourier approximation results: Faster implementations and stronger guarantees*, Numer. Algorithms, 63 (2013), pp. 239–263.

[52] Y. SHI AND A. ANANDKUMAR, *Higher-Order Count Sketch: Dimensionality Reduction That Retains Efficient Tensor Operations*, preprint, https://arxiv.org/abs/1901.11261, 2019.

[53] Y. SUN, Y. GUO, J. A. TROPP, AND M. UDELL, *Tensor random projection for low memory dimension reduction*, in Proceedings of the 32nd Conference on Neural Information Processing Systems Workshop on Relational Representation Learning, Montréal, Canada, 2018.

[54] G. TSITSIKAS AND E. E. PAPALEXAKIS, *The Core Consistency of a Compressed Tensor*, preprint, https://arxiv.org/abs/1811.07428, 2018.

[55] L. R. TUCKER, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[56] N. VANNIEUWENHOVEN, R. VANDEBRIL, AND K. MEERBERGEN, *A new truncation strategy for the higher-order singular value decomposition*, SIAM J. Sci. Comput., 34 (2012), pp. A1027–A1052, https://doi.org/10.1137/110836067.

[57] M. A. O. VASILESCU AND D. TERZOPOULOS, *Multilinear independent components analysis*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, IEEE, 2005, pp. 547–553.

[58] R. VERSHYNIN, *High-dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Ser. Statist. Probab. Math. 47, Cambridge University Press, 2018.

[59] R. VERSHYNIN, *Concentration Inequalities for Random Tensors*, preprint, https://arxiv.org/abs/1905.00802, 2019.

[60] Y. WANG, H.-Y. TUNG, A. J. SMOLA, AND A. ANANDKUMAR, *Fast and guaranteed tensor decomposition via sketching*, in NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, Montréal, Canada, Vol. 1, 2015, pp. 991–999.

[61] A. ZARE, A. OZDEMIR, M. A. IWEN, AND S. AVIYENTE, *Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor PCA*, Proc. IEEE, 106 (2018), pp. 1341–1358.

[62] H. ZHANG, W. HE, L. ZHANG, H. SHEN, AND Q. YUAN, *Hyperspectral image restoration using low-rank matrix recovery*, IEEE Trans. Geosci. Remote Sensing, 52 (2013), pp. 4729–4743.