Appl. Comput. Harmon. Anal. 66 (2023) 161–192

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha

Modewise operators, the tensor restricted isometry property, and low-rank tensor recovery ☆

Cullen A. Haselby [a], Mark A. Iwen [a], Deanna Needell [b], Michael Perlmutter [b,*], Elizaveta Rebrova [c]

[a] *Michigan State University, Department of Mathematics, and the Department of Computational Mathematics, Science and Engineering (CMSE), United States of America*
[b] *University of California, Los Angeles, Department of Mathematics, United States of America*
[c] *Princeton University, Department of Operations Research and Financial Engineering, United States of America*

ARTICLE INFO

Article history:
Received 14 October 2021
Received in revised form 16 August 2022
Accepted 27 April 2023
Available online 9 May 2023
Communicated by Ozgur Yilmaz

Keywords:
Tensor-structured data
Dimension reduction
Modewise measurements

ABSTRACT

Recovery of sparse vectors and low-rank matrices from a small number of linear measurements is well-known to be possible under various model assumptions on the measurements. The key requirement on the measurement matrices is typically the restricted isometry property, that is, approximate orthonormality when acting on the subspace to be recovered. Among the most widely used random matrix measurement models are (a) independent subgaussian models and (b) randomized Fourier-based models, allowing for the efficient computation of the measurements. For the now ubiquitous tensor data, direct application of the known recovery algorithms to the vectorized or matricized tensor is memory-heavy because of the huge measurement matrices to be constructed and stored. In this paper, we propose modewise measurement schemes based on subgaussian and randomized Fourier measurements. These modewise operators act on the pairs or other small subsets of the tensor modes separately. They require significantly less memory than the measurements working on the vectorized tensor, provably satisfy the tensor restricted isometry property and experimentally can recover the tensor data from fewer measurements and do not require impractical storage.

1. Introduction and prior work

Geometry preserving dimension reduction has become important in a wide variety of applications in the last two decades due to improved sensing capabilities and the increasing prevalence of massive data

☆ Mark A. Iwen partially supported by NSF DMS 1912706. Deanna Needell partially supported by NSF DMS 2011140. Mark A. Iwen, Deanna Needell, and Elizaveta Rebrova partially supported by NSF DMS 2106472.

* Corresponding author.
E-mail addresses: haselbyc@msu.edu (C.A. Haselby), markiwen@math.msu.edu (M.A. Iwen), deanna@math.ucla.edu (D. Needell), perlmutter@math.ucla.edu (M. Perlmutter), elre@princeton.edu (E. Rebrova).

https://doi.org/10.1016/j.acha.2023.04.007
1063-5203/

sets. This is motivated in part by the fact that the data one collects often consists of high-dimensional representations of intrinsically simpler and effectively lower-dimensional data. In such settings, randomized linear projections have been demonstrated to preserve the intrinsic geometric structure of the collected data in a wide range of applications in both computer science (where one often deals with finite data sets [14,1]) and signal processing (where manifold [4] and sparsity [17] assumptions are common). In this context, the vast majority of prior work has been focused on recovering vector data taking values in a set $S \subset \mathbb{R}^n$ using random linear maps into $\mathbb{R}^m$ with $m \ll n$ which are guaranteed to approximately preserve the norms of all elements in S. The focus of this paper is extending this line of work to higher-order tensors taking values in $\mathbb{R}^{n_1 \times \ldots \times n_d}$.

In the vector case, uniform guarantees for the approximate norm preservation for all sparse vectors, in the form of the restricted isometry property (RIP), have numerous applications. They include recovery algorithms that reconstruct all sparse vectors from a few linear measurements (such as, l_1-minimization [12,16,30], orthogonal matching pursuit [42], CoSaMP [31,16], iterative hard thresholding [8] and hard thresholding pursuit [15]). Extending these algorithms from sparse vector recovery to low-rank matrix or low-rank tensor recovery is very natural. Indeed, rank-r matrices (i.e., two-mode tensors) in $\mathbb{R}^{n \times n}$ can be recovered from $\mathcal{O}(rn)$ linear measurements [11,17]. Extensions to the low-rank higher-order tensor setting, however, are less straightforward due to, e.g., the more complicated structure of higher-order singular value decomposition and non-unique definition of the tensor rank. Still, there are many applications that motivate the use of tensors, ranging from video and longitudinal imaging [26,6] to machine learning [36,39] and differential equations [5,27]. Thus, while tensor applications are ubiquitous and moreover the tensors arising in these applications are extremely large-scale, few methods exist that do satisfactory tensor dimension reduction. Our goal here is thus to demonstrate a tensor dimension reduction technique that is computationally feasible (in terms of application and storage) and that guarantees preservation of geometry. As a motivating example, we consider the problem of tensor reconstruction from such dimension reduction measurements, and in particular the Tensor Iterative Hard Thresholding method is used for this purpose herein.

In [34], the authors propose tensor extensions of the Iterative Hard Thresholding (IHT) method for several tensor decomposition formats, namely the higher-order singular value decomposition (HOSVD), the tensor train format, and the general hierarchical Tucker decomposition. Additionally, the recent papers [20,19] extend the Tensor IHT method (TIHT) to low Canonical Polyadic (CP) rank and low Tucker rank tensors, respectively. TIHT as the name suggests is an iterative method that consists of one step that applies the adjoint of the measurement operator to the remaining residual and a second step that thresholds that signal proxy to a low-rank tensor. This method has seen provable guarantees for reconstruction under various geometry preserving assumptions on the measurement maps [34,20,19]. All these works however propose first reshaping a d-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ into an $\prod_{i=1}^{d} n_i$-dimensional vector $\mathbf{x}$ and then multiplying by an $m \times \prod_{i=1}^{d} n_i$ matrix A. Unfortunately, this means that the matrix A must be even larger than the original tensors $\mathcal{X}$. The main goal of this paper is to propose a more memory-efficient alternative to this approach.

In particular, we propose a *modewise* framework for low-rank tensor recovery. A general two-stage modewise linear operator $\mathcal{L} : \mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{R}^{m_1 \times \cdots \times m_{d'}}$ takes the form

$$\mathcal{L}(\mathcal{X}) := \mathcal{R}_2 \left(\mathcal{R}_1(\mathcal{X}) \times_1 A_1 \cdots \times_{\tilde{d}} A_{\tilde{d}} \right) \times_1 B_1 \cdots \times_{d'} B_{d'}, \tag{1}$$

where (i) $\mathcal{R}_1$ is a reshaping operator which reorganizes an $\mathbb{R}^{n_1 \times \cdots \times n_d}$ tensor into an $\mathbb{R}^{\tilde{n}_1 \times \cdots \times \tilde{n}_{\tilde{d}}}$ tensor, after which (ii) each $A_j \in \mathbb{R}^{\tilde{m}_j \times \tilde{n}_j}$ is applied to the resphaped tensor for $j = 1, \ldots, \tilde{d}$ via a modewise product (reviewed in Section 2), followed by (iii) an additional reshaping via $\mathcal{R}_2$ into an $\mathbb{R}^{m'_1 \times \cdots \times m'_{d'}}$ tensor, and finally (iv) additional j-mode products with the matrices $B_j \in \mathbb{R}^{m_j \times m'_j}$ for $j = 1, \ldots, d'$. More general n-stage modewise operators can be defined similarly. First analyzed in [21,23] for aiding in the rapid computation of the CP decomposition, such modewise compression operators offer a wide variety of

computational advantages over standard vector-based approaches (in which $\mathcal{R}_1$ is a vectorization operator so that $\tilde{d} = 1$, $A_1 = A \in \mathbb{R}^{m \times \prod_j^d n_j}$ is a standard Johnson-Lindenstrauss map (see, e.g., [29]), and all remaining operators $\mathcal{R}_2, B_1, \dots$ are the identity). In particular, when $\mathcal{R}_1$ is a more modest reshaping (or even the identity) the resulting modewise linear transforms can be formed using significantly fewer random variables (effectively, independent random bits), and stored using less memory by avoiding the use of a single massive $m \times \prod_j^d n_j$ matrix. In addition, such modewise linear operators also offer trivially parallelizable operations, faster serial data evaluations than standard vectorized approaches do for structured data (see, e.g., [23]), and the ability to better respect the multimodal structure of the given tensor data.

Related Work: This paper is directly motivated by recent work on low-rank tensor recovery using vectorized measurements [34,20,19]. In particular, we consider the same class of low-rank tensors as in [34], but utilize modewise compression maps rather than a purely vectorization based approach. We also note other recent work involving the analysis of modewise maps for tensor data include, e.g., applications in kernel learning methods which effectively use modewise operators specialized to finite sets of rank-one tensors [2], as well as a variety of works in the computer science literature aimed at compressing finite sets of low-rank (with respect to, e.g., CP and tensor train decompositions [33]) tensors. Additionally, we note more general results involving extensions of bounded orthonormal sampling results to the tensor setting [23,3] which apply to finite sets of arbitrary tensors.

Contributions: The purpose of this paper is to use the framework of modewise measurement operators (see Equation (1)) to create highly structured and computationally efficient measurement maps. We aim to provide both theoretical guarantees and empirical evidence that several of these modewise maps allow for the efficient recovery of tensors with low-rank HOSVD decompositions. This represents the first study of such modewise maps for performing norm-preserving dimension reduction of nontrivial infinite sets of elements in (tensorized) Euclidean spaces, and so provides a general framework for generalizing the use of such maps to other types of, e.g., low-rank tensor models. Our main theoretical results are presented in Section 3. In particular, Theorem 3.3 provides a sufficient condition for a modewise map to satisfy a TRIP. Theorem 3.8 then provides sufficient considitions for a two-step map which combines modewise maps and a vectorization based approach to also satisfy TRIP. Corollaries 3.4, 3.6, 3.9 and 3.11 then provide examples of maps, constructed from either subgaussian matrices or subsampled orthogonal matrices with random sign that will satisfy these assumptions with high probability. While previous work has shown that it is possible to construct norm-preserving modewise embeddings of either finite sets [23] or low-dimensional subspaces (see, e.g., [21,28]), this is the first work to extend these techniques to the much larger set of all tensors with a low-rank HOSVD decomposition in order to obtain modewise embeddings with the Tensor Restricted Isometry Property (TRIP). Additionally, having obtained modewise TRIP operators, we then consider low-rank tensor recovery via Tensor IHT (TIHT). We also provide an empirical demonstration of the good performance such modewise maps can provide for tensor recovery in Section 4.1.

A Motivation for Low Memory Measurements: One example in which low storage requirements for linear measurements are particularly valuable is the use streaming algorithms for tensor reconstruction in the big data setting (see, e.g., [37]). In such settings one does not have access to the entire large tensor one wishes to approximate all at once, but instead receives the tensor over an extended period of time via a series of updates. For example, suppose that one is aggregating tensor data over time via the simple additive rule

$$\mathcal{X}_t = \mathcal{X}_{t-1} + \Delta\mathcal{X}_t, \qquad \mathcal{X}_0 = \mathcal{0}$$

based on updates $\Delta\mathcal{X}_t$. Then, at a later time, one needs to reconstruct an estimate of $\mathcal{X}_T = \sum_{t=1}^{T} \Delta\mathcal{X}_t \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ for some $T > 0$ large. Here the simple solution of simply storing every intermediate tensor $\mathcal{X}_t$ in uncompressed form is unattractive when the overall tensor dimensions are large because it requires a huge storage commitment over a large number of updates for simply aggregating the final tensor $\mathcal{X}_T$. Alternatively, one may reduce the storage costs by instead storing small linear sketches of the intermediate

tensors. In this setting a linear operator $\mathcal{L} : \mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{R}^{m_1 \times \cdots \times m_{d'}}$ with $\prod_{\ell=1}^{d'} m_\ell \ll \prod_{\ell=1}^{d} n_\ell$ can be used to iteratively update the much smaller sketch

$$\mathcal{L}(\mathcal{X}_t) = \mathcal{L}(\mathcal{X}_{t-1}) + \mathcal{L}(\Delta\mathcal{X}_t)$$

over the course of the many updates $t = 0, \ldots, T$. The final sketch $\mathcal{L}(\mathcal{X}_T)$ can then be used to approximately recover $\mathcal{X}_T$ at a later date in a one-time memory-expensive application of, e.g., TIHT.

A potential drawback of this approach is that one must store $\mathcal{L}$ itself in addition to storing the $\mathcal{L}(\mathcal{X}_{t-1})$ and $\mathcal{L}(\Delta\mathcal{X}_t)$. Therefore, in order for this method to be useful, the cost of storing $\mathcal{L}$ must be less than the savings accrued by not storing the $\mathcal{X}_t$. Indeed, even in the traditional compression setting, one needs to store the measurement operator itself and without a low memory approach, that storage itself will often exceed any savings gained by the compression of the data. Hence, with such motivating applications in mind, we focus on developing low-memory tensor sketches $\mathcal{L}$.

Paper Outline: The rest of this paper is organized as follows. In Section 2, we will provide a brief review of basic tensor definitions. In Section 3, we will state our main results, which we then prove in Section 5. In Section 4, we discuss applications of our results recovering low-rank tensors via the TIHT, and present numerical results. In Section 6, we provide a short conclusion and discussion of directions for future work. Proofs of auxiliary results are provided in the appendices.

2. Tensor prerequisites

In this section, we briefly review some basic definitions concerning tensors. For further overview, we refer the reader to [24]. Let $d \geqslant 1$, $n_1, \ldots, n_d \geqslant 1$ be integers, and $[n_j] := \{1, \ldots, n_j\}$ for all $j = 1, \ldots, d$. For a multi-index $\mathbf{i} = (i_1, \ldots, i_d) \in [n_1] \times \cdots \times [n_d]$, we will denote the $\mathbf{i}$-th entry of a d-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ by $\mathcal{X}_\mathbf{i}$. When convenient we will also denote the entries by $\mathcal{X}(i_1, \ldots, i_d)$, $\mathcal{X}_{i_1,\ldots,i_d}$, or $\mathcal{X}(\mathbf{i})$. For the remainder of this work, we will use bold text to denote vectors (i.e., one-mode tensors), capital letters to denote matrices (i.e., two-mode tensors) and use calligraphic text for all other tensors.

2.1. Modewise multiplication and j-mode products:

For $1 \leqslant j \leqslant d$, the *$j$-mode product* of d-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_{j-1} \times n_j \times n_{j+1} \times \cdots \times n_d}$ with a matrix $U \in \mathbb{R}^{m_j \times n_j}$ is another d-mode tensor $\mathcal{X} \times_j U \in \mathbb{R}^{n_1 \times \cdots \times n_{j-1} \times m_j \times n_{j+1} \times \cdots \times n_d}$. Its entries are given by

$$(\mathcal{X} \times_j U)_{i_1,\ldots,i_{j-1},\ell,i_{j+1},\ldots,i_d} = \sum_{i_j=1}^{n_j} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} U_{\ell,i_j} \tag{2}$$

for all $(i_1, \ldots, i_{j-1}, \ell, i_{j+1}, \ldots, i_d) \in [n_1] \times \cdots \times [n_{j-1}] \times [m_j] \times [n_{j+1}] \times \cdots \times [n_d]$. If $\mathbf{y}_1, \ldots \mathbf{y}_d$ are vectors, $\mathbf{y}_j \in \mathbb{R}^{n_j}$, we define their outer product $\mathcal{Y} = \bigcirc_{j=1}^{d} \mathbf{y}_j \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ to be the d-mode tensor in $\mathbb{R}^{n_1 \times \cdots \times n_d}$ whose entries are given by

$$\mathcal{Y}_{i_1,\ldots,i_d} = \prod_{j=1}^{d} (\mathbf{y}_j)_{i_j}.$$

In particular, if $\mathcal{Y}$ has this form, then one may use (2) to see that

$$\mathcal{Y} \times_j U = \left(\bigcirc_{i=1}^{d} \mathbf{y}_i \right) \times_j U = \left(\bigcirc_{i=1}^{j-1} \mathbf{y}_i \right) \bigcirc U\mathbf{y}_j \bigcirc \left(\bigcirc_{i=j+1}^{d} \mathbf{y}_i \right). \tag{3}$$

For further discussion of the properties of modewise products, please see [21,24].

2.2. HOSVD decomposition and multilinear rank

Let $\mathbf{r} = (r_1, \dots, r_d)$ be a multi-index in $[n_1] \times \cdots \times [n_d]$. We say that a d-mode tensor $\mathcal{X}$ has **multilinear rank** or **HOSVD rank** at most $\mathbf{r}$ if there exist subspaces $\mathcal{U}_1 \subset \mathbb{R}^{n_1}, \dots, \mathcal{U}_d \subset \mathbb{R}^{n_d}$ such that

$$\dim \mathcal{U}_i = r_i \quad \text{and} \quad \mathcal{X} \in \bigotimes_{i=1}^{d} \mathcal{U}_i,$$

where $\bigotimes_{i=1}^{d} \mathcal{U}_i$ denotes the tensor product of the subspaces $\mathcal{U}_i$ here. We note that a tensor $\mathcal{X}$ has rank at most $\mathbf{r} = (r_1, \dots, r_d)$ if and only if there exists a core tensor $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ such that

$$\mathcal{X} = \mathcal{C} \times_1 U^1 \times_2 \dots \times_d U^d = \sum_{k_d=1}^{r_d} \dots \sum_{k_1=1}^{r_1} \mathcal{C}(k_1, \dots k_d) \bigcirc_{i=1}^{d} \mathbf{u}^i_{k_i}, \tag{4}$$

where, for each $1 \leqslant i \leqslant d$, $\mathbf{u}^i_1, \dots, \mathbf{u}^i_{r_i}$ is an orthonormal basis for $\mathcal{U}_i$, and U^i is the $n_i \times r_i$ matrix $U^i = (\mathbf{u}^i_1, \dots, \mathbf{u}^i_{r_i})$. A factorization of the form (4) is called a **Higher-Order Singular Value Decomposition (HOSVD)** of the tensor $\mathcal{X}$. It is well-known (see e.g., [34]) that we may assume that the core tensor $\mathcal{C}$ has orthogonal subtensors in the sense that for all $1 \leqslant i \leqslant d$, we have $\langle \mathcal{C}_{k_i=p}, \mathcal{C}_{k_i=q} \rangle = 0$ for all $p \neq q$, where $\mathcal{C}_{k_i=p}$ is $(d-1)$-mode subtensor of $\mathcal{C}$ only containing entries where $k_i = p$, i.e.

$$\sum_{\substack{k_j=1 \\ j \neq i}}^{r_j} \mathcal{C}(k_1, \dots, p, \dots, k_d)\mathcal{C}(k_1, \dots, q, \dots, k_d) = 0 \quad \text{unless } p = q. \tag{5}$$

We also note that, since each of the $\{\mathbf{u}^i_{k_i}\}_{i=1}^{r_i}$ form an orthonormal basis for $\mathcal{U}_i$, we have $\|\mathcal{C}\|_F = \|\mathcal{X}\|_F := \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$, where here $\langle \cdot, \cdot \rangle$ denotes the trace inner product.

Remark 2.1. The **Canonical Polyadic (CP) rank** of a d-mode tensor is the minimum number of rank-one tensors (i.e., outer products of d vectors) required to represent the tensor as a sum. If $\mathcal{X}$ has HOSVD rank $\mathbf{r}$ then (4) implies $\mathcal{X}$ has CP rank at most $\Pi_{i=1}^{d} r_i$. (In particular, if $r_i = r$ for all i, then $\mathcal{X}$ has CP rank at most r^d.)

2.3. Restricted isometry properties and tensors

Definition 2.2. [RIP$(\varepsilon, \mathcal{S})$ property] We say that a linear map $\mathcal{A}$, defined on a normed vector space with norm $\|\cdot\|$, has the RIP$(\varepsilon, \mathcal{S})$ property if for all elements $s \in \mathcal{S}$

$$(1-\varepsilon)\|s\|^2 \leqslant \|\mathcal{A}(s)\|^2 \leqslant (1+\varepsilon)\|s\|^2 \tag{6}$$

We emphasize that the set $\mathcal{S}$ in Definition 2.2 can be a subset of any normed vector space (not necessarily a tensor space). In the following definition, we will focus on the generalized Frobenius norm which we define by

$$\|\mathcal{X}\|_F^2 := \sum_{i_d=1}^{n_d} \dots \sum_{i_1=1}^{n_1} |\mathcal{X}(i_1, \dots i_d)|^2.$$

Definition 2.3. [TRIP$(\delta, \mathbf{r})$ property] We say that a linear map $\mathcal{A}$ has the TRIP$(\delta, \mathbf{r})$ property if for all $\mathcal{X}$ with HOSVD rank at most $\mathbf{r}$ we have

$$(1-\delta)\|\mathcal{X}\|_F^2 \leqslant \|\mathcal{A}(\mathcal{X})\|_F^2 \leqslant (1+\delta)\|\mathcal{X}\|_F^2 \tag{7}$$

2.4. Reshaping and the HOSVD

For simplicity we will assume below, and for the rest of this paper, that there exist $n, r \geqslant 1$ such that $n_i = n, r_i = r$ for $1 \leqslant i \leqslant d$. We note that this assumption is made only for the sake of clarity, and all of our analysis can be extended to the general case.

We let κ be an integer which divides d and let $d' := d/\kappa$. Consider the *reshaping operator*

$$\mathcal{R} : \bigotimes_{i=1}^{d} \mathbb{R}^n \to \bigotimes_{i=1}^{d'} \mathbb{R}^{n^\kappa}$$

that flattens every κ modes of a tensor into one. Note that $\mathcal{R}$ decreases the total number of modes from d to $d' = d/\kappa$. Formally, $\mathcal{R}$ is defined to be the unique linear operator such that on rank-one tensors it acts as

$$\mathcal{R}\left(\bigcirc_{i=1}^{d} \mathbf{x}^i\right) := \bigcirc_{i=1}^{d'} \left(\bigotimes_{\ell=1+\kappa(i-1)}^{\kappa i} \mathbf{x}^\ell \right) =: \bigcirc_{i=1}^{d'} \mathring{\mathbf{x}}^i,$$

where $\otimes$ denotes the Kronecker product when applied to vectors. We observe that if a tensor $\mathcal{X}$ has a form (4), then its reshaping $\mathring{\mathcal{X}} := \mathcal{R}(\mathcal{X})$ is the d'-mode tensor $\mathring{\mathcal{X}} \in \bigotimes_{i=1}^{d'} \mathbb{R}^{n^\kappa}$ with HOSVD rank at most $\mathbf{r}' := (r^\kappa, \dots, r^\kappa)$ given by

$$\mathring{\mathcal{X}} = \sum_{j_{d'}=1}^{r^\kappa} \dots \sum_{j_1=1}^{r^\kappa} \mathring{\mathcal{C}}(j_1, \dots, j_{d'}) \bigcirc_{\ell=1}^{d'} \mathring{\mathbf{u}}^\ell_{j_\ell}, \tag{8}$$

where the new component vectors $\mathring{\mathbf{u}}^\ell_{j_\ell}$ are obtained by taking Kronecker product of the appropriate $\mathbf{u}^i_{k_i}$, and where $\mathring{\mathcal{C}} \in \mathbb{R}^{r^\kappa \times \dots \times r^\kappa}$ is a reshaped version of $\mathcal{C}$ from (4). Since each of the $\{\mathbf{u}^i_{k_i}\}_{k_i=1}^{r}$ was an orthonormal basis for $\mathcal{U}_i$, it follows that $\{\mathring{\mathbf{u}}^i_{j_i}\}_{j_i=1}^{r^\kappa}$ is an orthonormal basis for $\mathring{\mathcal{U}}_i := \operatorname{span}\left(\{\mathring{\mathbf{u}}^i_{j_i}\}_{j_i=1}^{r^\kappa}\right)$.

3. Main results: modewise TRIP

For $1 \leqslant i \leqslant d'$, let A_i be an $m \times n^\kappa$ matrix, let $\mathcal{A} : \mathbb{R}^{n^\kappa \times \dots \times n^\kappa} \to \mathbb{R}^{m \times \dots \times m}$ be the linear map which acts modewise on d'-mode tensors by

$$\mathcal{A}(\mathcal{Y}) = \mathcal{Y} \times_1 A_1 \times_2 \dots \times_{d'} A_{d'}. \tag{9}$$

Let $\mathcal{X}$ be a d mode tensor with HOSVD decomposition given by (4). By (3) and (8), we have that

$$\mathcal{A}(\mathcal{R}(\mathcal{X})) = \mathcal{A}(\mathring{\mathcal{X}}) = \sum_{j_{d'}=1}^{r^\kappa} \dots \sum_{j_1=1}^{r^\kappa} \mathring{C}(j_1, \dots, j_{d'})(A_1 \mathring{\mathbf{u}}^1_{j_1} \circ \dots \circ A_{d'} \mathring{\mathbf{u}}^{d'}_{j_{d'}}). \tag{10}$$

Our first main result will show that $\mathcal{A}$ satisfies the TRIP$(\delta, \mathbf{r})$ property under the assumption that each of the A_i satisfies a restricted isometry property on the set $\mathcal{S}_{1,2}$ defined below, which corresponds to all unit norm tensors that either have rank $(1, \dots, 1)$ or can be written as the sum of two tensors with rank $(1, \dots, 1)$.

Definition 3.1. [The set $\mathcal{S}_{1,2}$] Consider a set of vectors in $\mathbb{R}^{n^\kappa}$,

$$S_1 := \{\mathring{\mathbf{u}} \mid \mathring{\mathbf{u}} = \otimes_1^\kappa \mathbf{u}^i, \mathbf{u}^i \in \mathbb{S}^{n-1}\}, \tag{11}$$

and let $\mathcal{S}_2 := \left\{ \frac{\mathbf{x}+\mathbf{y}}{\|\mathbf{x}+\mathbf{y}\|_2} \;\middle|\; \mathbf{x},\mathbf{y} \in \mathcal{S}_1 \text{ s.t. } \langle \mathbf{x},\mathbf{y}\rangle = 0 \right\}$. For the rest of this text we will let $\mathcal{S}_{1,2} := \mathcal{S}_1 \cup \mathcal{S}_2$, and note that $\mathcal{S}_{1,2} \subseteq \mathbb{S}^{n^\kappa - 1}$.

More precisely, will show that $\mathcal{A} \circ \mathcal{R}$ satisfies the TRIP$(\delta, \mathbf{r})$ property under the assumption that each of the A_i satisfy the RIP$(\varepsilon, \mathcal{S}_{1,2})$, where ε is a suitably chosen parameter depending on δ. In the case where $\mathbf{r} = \mathbf{1} := (1, 1, \ldots, 1)$, this is nearly trivial. Indeed, if $\mathring{\mathcal{X}} = c \bigcirc_{\ell=1}^{d'} \mathring{\mathbf{u}}^\ell$, and $\mathcal{A}$ is the map defined in (9), then we have

$$\left\| \mathcal{A}(\mathring{\mathcal{X}}) \right\| = \left\| c \bigcirc_{\ell=1}^{d'} A_i \mathring{\mathbf{u}}^\ell \right\| = |c| \prod_{\ell=1}^{d} \|A_i \mathring{\mathbf{u}}^\ell\|.$$

Therefore, since $\left\| \mathring{\mathcal{X}} \right\| = |c| \prod_{\ell=1}^{d} \|\mathring{\mathbf{u}}^\ell\|$, we immediately obtain the following proposition.

Proposition 3.2. *Suppose that $\mathcal{A}$ is defined as per (9) and that each of the A_i have RIP$(\varepsilon, \mathcal{S}_{1,2})$ property. Let $\delta = \max\{(1+\varepsilon)^d - 1, 1 - (1-\varepsilon)^d\}$ and assume that $\delta < 1$. Then $\mathcal{A} \circ \mathcal{R}$ satisfies the TRIP$(\delta, \mathbf{1})$ property, that is,*

$$(1-\delta)\|\mathcal{X}\|^2 \leqslant \|\mathcal{A}(\mathcal{R}(\mathcal{X}))\|^2 \leqslant (1+\delta)\|\mathcal{X}\|^2$$

for all $\mathcal{X}$ with HOSVD rank $\mathbf{1} = (1, 1, \ldots, 1)$.

Our first main result is the following theorem which is the analogue for Proposition 3.2 for $r \geqslant 2$. It shows that if each of the A_i satisfies RIP$(\varepsilon, \mathcal{S}_{1,2})$ property for a suitable value of ε, then $\mathcal{A}$ has the TRIP$(\delta, \mathbf{r})$ property.

Theorem 3.3. *Suppose that $\mathcal{A}$ is defined as per (9) and that each of the A_i have RIP$(\varepsilon, \mathcal{S}_{1,2})$ property. Let $r \geqslant 2$, let $\delta = 4d' r^d \varepsilon$ and assume that $\delta < 1$. Then $\mathcal{A} \circ \mathcal{R}$ satisfies the TRIP$(\delta, \mathbf{r})$ property, i.e.,*

$$(1-\delta)\|\mathcal{X}\|^2 \leqslant \|\mathcal{A}(\mathcal{R}(\mathcal{X}))\|^2 \leqslant (1+\delta)\|\mathcal{X}\|^2 \tag{12}$$

for all $\mathcal{X}$ with HOSVD rank less than $\mathbf{r} = (r, r, \ldots, r)$.

Proof. See Section 5.3. □

The following corollary shows that we may pick the matrices A_i to have i.i.d. properly normalized subgaussian entries.

Corollary 3.4. *Let $r \geqslant 2$, and let $\mathbf{r} = (r, r, \ldots, r)$. Suppose that $\mathcal{A}$ is defined as per (9) and that each of the $A_i \in \mathbb{R}^{m \times n^\kappa}$ has i.i.d. $\frac{1}{m}$-subgaussian entries, for all $i = 1, \ldots, d'$, where $d' = d/\kappa$ for $\kappa \geqslant 2$, and suppose that $0 < \eta, \delta < 1$. Let*

$$m \geqslant C\delta^{-2} r^{2d} \max\left\{ \frac{nd^2 \ln(\kappa)}{\kappa}, \frac{d^2}{\kappa^2} \ln\left(\frac{d}{\kappa\eta}\right) \right\} \tag{13}$$

for a sufficiently large constant C. Then $\mathcal{A} \circ \mathcal{R}$ satisfies TRIP$(\delta, \mathbf{r})$ property (7) with probability at least $1 - \eta$.

Proof. See Section 5.3. □

If one compares Corollary 3.4 to the analogous result in [34], they may note that our bounds depend on r^{2d} rather than r^d. However, in many applications of interest the total dimension $\prod_{i=1}^d n_i$ is orders of magnitude larger than the number of modes d. For example, a five-minute 1080p video shot at 24 frames per second is naturally modeled as a tensor in $\mathbb{R}^{1920\times 1080\times 3\times 7200}$, in which case we have $d = 4$ and $\prod_{i=1}^4 n_i \approx 4.48 \times 10^{10}$. Therefore, in this paper, we will focus on the dependence in on the total dimension rather than the number of modes.

For another possible choice of the A_i, we consider the set of Subsampled Orthogonal with Random Sign matrices defined below. Note, in particular, that this class includes subsampled Fourier (i.e., discrete cosine and sine) matrices.

Definition 3.5 *(Subsampled Orthogonal with Random Sign (SORS) matrices).* Let $F \in \mathbb{R}^{n\times n}$ denote an orthonormal matrix obeying

$$F^*F = I \quad \text{and} \quad \max_{i,j} |F_{ij}| \leq \frac{\Delta}{\sqrt{n}} \tag{14}$$

for some $\Delta > 0$. Let $H \in \mathbb{R}^{m\times n}$ be a matrix whose rows are chosen i.i.d. uniformly at random from the rows of F. We define a Subsampled Orthogonal with Random Sign (SORS) measurement ensemble as $A = \sqrt{\frac{n}{m}} HD$, where $D \in \mathbb{R}^{n\times n}$ is a random diagonal matrix whose the diagonal entries are i.i.d. ± 1 with equal probability.

Analogous to Corollary 3.4, the following result shows that we may choose our matrices A_i to be SORS matrices in Theorem 3.3.

Corollary 3.6. *Let $r \geqslant 2$ and let $\mathbf{r} = (r, r, \ldots, r)$. Suppose that $\mathcal{A}$ is defined as per (9) and that each of the $A_i \in \mathbb{R}^{m\times n^\kappa}$ is a SORS matrix with $\Delta \leqslant C'$ for a universal constant C', as per Definition 3.5, for all $i = 1, \ldots, d'$, where $d' = d/\kappa$ for $\kappa \geqslant 2$. Furthermore, suppose that $0 < \eta, \delta < 1$. Let*

$$m \geqslant C_1 \delta^{-2} r^{2d} \frac{nd^2 \ln(\kappa)}{\kappa} \cdot L, \tag{15}$$

where

$$L = \ln\left(\frac{2d}{\kappa\eta}\right) \ln\left(\frac{2en^\kappa d}{\kappa\eta}\right) \ln^2\left[C_2 \delta^{-2} r^{2d} \frac{nd^2 \ln(\kappa)}{\kappa} \ln\left(\frac{2d}{\kappa\eta}\right)\right] \tag{16}$$

and C_1, C_2 are sufficiently large absolute constants. Then $\mathcal{A} \circ \mathcal{R}$ satisfies TRIP($\delta, \mathbf{r}$) property (7) with probability at least $1 - \eta$.

Proof. See Section 5.3. □

If we wish, to further improve embedding dimension of $m^{d'}$ provided by Corollaries 3.4 and 3.6, we can apply a secondary compression, analogous to the one used in [21], by letting

$$\mathcal{A}_{2nd}(\mathcal{X}) := A_{2\text{nd}}(\text{vect}(\mathcal{A}(\mathcal{R}(\mathcal{X})))), \tag{17}$$

where $\mathcal{A}$ and $\mathcal{R}$ are as in Theorem 3.3, vect is a vectorization operator which reshapes a d-mode tensor into a vector by lexicographic ordering, and $A_{2\text{nd}}$ is a matrix which satisfies an RIP property on the range of vect$\circ\mathcal{A}\circ\mathcal{R}$. Notably, we will show that the two-step map, $\mathcal{A}_{2nd}$ is able to achieve greater dimension reduction than its one-step counterpart $\mathcal{A} \circ \mathcal{R}$. While it is true that $\mathcal{A}_{2nd}$ has a higher memory cost than $\mathcal{A} \circ \mathcal{R}$, it has lower memory cost than a vectorization based approach as we shall discuss further in Remark 3.13.

In Theorem 3.8 stated below, we will show that $\mathcal{A}_{2nd}$ satisfies $TRIP(\delta, \mathbf{r})$ for suitably chose parameters. One of the key challenges in doing this is that, for any given i, the new factor vectors $\{A_i \mathring{u}^i_{j_i}\}_{j_i=1}^{r^\kappa}$ defined as in (10) are no longer orthogonal to one another. Therefore (10) is not an HOSVD decomposition of $\mathcal{A}(\mathring{\mathcal{X}})$, and the HOSVD rank of $\mathcal{A}(\mathring{\mathcal{X}})$ might be much larger than the HOSVD rank of $\mathring{\mathcal{X}}$. However, one may overcome this difficulty by observing that, with high probability, $\mathcal{A}(\mathcal{R}(\mathcal{X}))$ will belong to the following set of nearly orthogonal tensors.

Definition 3.7 *(Nearly orthogonal tensors $\mathcal{B}_{R,\mu,\theta,\mathbf{r}}$).* Let $\mathcal{B}_{R,\mu,\theta,\mathbf{r}}$ be the set of d-mode tensors in $\mathcal{X} \in \mathbb{R}^{n\times\ldots\times n}$ that may be written in standard form (4) such that

(a) $\|\mathbf{u}^i_{k_i}\|_2 \leqslant R$ for all i and k_i,
(b) $|\langle \mathbf{u}^i_{k_i}, \mathbf{u}^i_{k'_i}\rangle| \leqslant \mu$ for all $k_i \neq k'_i$,
(c) the core tensor $\mathcal{C}$ satisfies $\|\mathcal{C}\|_F = 1$,
(d) $\mathcal{C}$ has orthogonal subtensors in the sense that (5) holds for all $1 \leqslant i \leqslant d$,
(e) $\|\mathcal{X}\|_F \geqslant \theta$.

Our next main result is the following theorem which shows that $\mathcal{A}_{2nd}$ satisfies $TRIP(\delta, \mathbf{r})$ for suitably chosen parameters.

Theorem 3.8. *Let $r \geqslant 2$ and let $\mathbf{r} = (r, r, \ldots, r) \in \mathbb{R}^d$. Suppose that $\mathcal{A}$ and $\mathcal{A}_{2nd}$ are defined as in (9) and (17). Let $d' = d/\kappa$ and assume that A_i satisfies the $RIP(\varepsilon, \mathcal{S}_{1,2})$ property for all $i = 1, \ldots, d'$, where $\delta = 12d'r^d\varepsilon < 1$. Assume that A_{2nd} satisfies the $RIP\big(\delta/3, \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}\big)$, property where $\mathbf{r}' = (r^\kappa, \ldots, r^\kappa) \in \mathbb{R}^{d'}$. Then, $\mathcal{A}_{2nd}$ will satisfy the $TRIP(\delta, \mathbf{r})$ property, i.e.,*

$$(1-\delta)\|\mathcal{X}\|^2 \leqslant \|\mathcal{A}_{2nd}(\mathcal{X})\|^2 \leqslant (1+\delta)\|\mathcal{X}\|^2$$

for all $\mathcal{X}$ with HOSVD rank at most $\mathbf{r}$.

Proof. See Section 5.4. □

The following two corollaries show that we may choose the matrices A_i and $A_{2\text{nd}}$ to be either $\frac{1}{m}$-subgaussian or SORS matrices. We also note that it is possible to produce other variants of these corollaries where, for example, one takes each A_i to be subgaussian and lets $A_{2\text{nd}}$ be a SORS matrix.

Corollary 3.9. *Let $r \geqslant 2$ and let $\mathbf{r} = (r, r, \ldots, r) \in \mathbb{R}^d$. Suppose that $\mathcal{A}$ and $\mathcal{A}_{2nd}$ are defined as in (9) and (17), and that all of the $A_i \in \mathbb{R}^{m\times n^\kappa}$ have i.i.d. $\frac{1}{m}$-subgaussian entries for all $i = 1, \ldots, d'$, where $d' = d/\kappa$, and suppose that $0 < \eta, \delta < 1$. Let*

$$m \geqslant C\delta^{-2} r^{2d} \max\left\{\frac{nd^2\ln(\kappa)}{\kappa}, \frac{d^2}{\kappa^2}\ln\left(\frac{2d}{\kappa\eta}\right)\right\}, \tag{18}$$

and let $A_{2nd} \in \mathbb{R}^{m_{2nd}\times m}$ be a $\frac{1}{m_{2nd}}$-subgaussian matrix with i.i.d. entries with

$$m_{2nd} \geqslant C\delta^{-2} \max\left\{\left(\frac{r^d\kappa + dmr^\kappa}{\kappa}\right)\ln\left(\frac{d}{\kappa}+1\right) + \frac{dmr^\kappa}{\kappa}\ln\left(1+\delta r^d\right) + \frac{d^2mr^\kappa\delta}{\kappa^2}, \ln\left(\frac{2}{\eta}\right)\right\}. \tag{19}$$

Then, $\mathcal{A}_{2nd}$ satisfies the $TRIP(\delta, \mathbf{r})$ property, i.e.,

$$(1-\delta)\|\mathcal{X}\|^2 \leqslant \|\mathcal{A}_{2nd}(\mathcal{X})\|^2 \leqslant (1+\delta)\|\mathcal{X}\|^2,$$

for all $\mathcal{X}$ with HOSVD rank at most $\mathbf{r}$ with probability at least $1-\eta$.

Proof. See Section 5.4. □

Remark 3.10. Note that applying the reshaping operator (with $\kappa > 1$) is necessary in order for us to actually achieve dimension reduction in the first step. Indeed, if $\kappa = 1$, then (18) requires $m > n^{\kappa}$. We also note that when other parameters are held fixed, the final dimension m_{2nd} will be required to be $\mathcal{O}(n)$, $\mathcal{O}(\delta^{-2})$, $\mathcal{O}(\ln(\eta^{-1}))$ or $\mathcal{O}(r^{2d})$. While the dependence on the number modes d is exponential, we are primarily interested in cases where n is large in comparison to the rank or the number of modes. In this case, the terms involving n will dominate the terms involving r^d. In Section 4.1, we will show that TRIP-dependent tensor recovery methods (e.g., tensor iterative hard thresholding, discussed in Section 4), successfully work for $d = 4$ and $\kappa = 2$.

In [34], the author considered i.i.d. subgaussian measurements applied to the vectorizations of low-rank tensors and proved that the $TRIP(\delta, \mathbf{r})$ property will hold with probability at least $1 - \eta$ if the target dimension satisfies

$$m_{final} \geqslant C\delta^{-2} \max\{(r^d + dnr) \ln d, \ln(\eta^{-1})\}.$$

We note this bound has the same computational complexity as ours with respect to n, δ, and η. While their result has much better dependence on r, here, we are primarily interested in high-dimensional, low-rank tensors and therefore are primarily concerned with the dependence on n. It is not known if the results from [34] are optimal, but it is believed that these results are likely almost optimal. Therefore, we believe that our results are likely almost optimal with respect to n, δ, and η as well.

Corollary 3.11. *Let $r \geqslant 2$ and let $\mathbf{r} = (r, r, \ldots, r)$. Suppose that $\mathcal{A}$ and $\mathcal{A}_{2nd}$ are defined as in (9) and (17), and that all of the $A_i \in \mathbb{R}^{m \times n^{\kappa}}$ are SORS matrices (as per Definition 3.5) for all $i = 1, \ldots, d'$, where $d' = d/\kappa$. Furthermore, suppose that $0 < \eta, \delta < 1$ and, as in (15), let*

$$m \geqslant C_1 \delta^{-2} r^{2d} \frac{nd^2 \ln(\kappa)}{\kappa} \cdot L, \text{ where } L \text{ is defined by (16).} \tag{20}$$

Next, let $A_{2nd} \in \mathbb{R}^{m_{2nd} \times m}$ also be a SORS matrix with

$$m_{2nd} \geqslant C\delta^{-2} \left[\frac{r^d \kappa + dmr^{\kappa}}{\kappa} \ln\left(\frac{d}{\kappa} + 1\right) + \frac{dmr^{\kappa}}{\kappa} \ln\left(1 + \delta r^d\right) + \frac{d^2 m r^{\kappa} \delta}{\kappa^2} \right] \cdot \tilde{L}, \tag{21}$$

where

$$\tilde{L} = \ln^2\left(\frac{c_1}{\delta^2}\left(\ln\frac{4}{\eta}\right)\left[\frac{r^d\kappa + dmr^{\kappa}}{\kappa} \ln\left(\frac{d}{\kappa} + 1\right) + \frac{dmr^{\kappa}}{\kappa} \ln\left(1 + \delta r^d\right) + \frac{d^2 m r^{\kappa}\delta}{\kappa^2}\right]\right) \ln\left(\frac{4}{\eta}\right) \ln\left(\frac{4em}{\eta}\right).$$

Then, $\mathcal{A}_{2nd}$ satisfies the TRIP($\delta, \mathbf{r}$) property, i.e.,

$$(1 - \delta)\|\mathcal{X}\|^2 \leqslant \|\mathcal{A}_{2nd}(\mathcal{X})\|^2 \leqslant (1 + \delta)\|\mathcal{X}\|^2$$

holds for all $\mathcal{X}$ with HOSVD rank at most $\mathbf{r}$ with probability at least $1 - \eta$.

Proof. See Section 5.4. □

Remark 3.12. Similar to the subgaussian case, we note that reshaping (with $\kappa > 1$) is needed in order for us to achieve dimension reduction in the first compression. We also note that the final dimension is $\mathcal{O}(n\text{polylog}(n))$, $\mathcal{O}(\delta^{-2}\text{polylog}(\delta^{-2}))$, $\text{polylog}(\eta^{-1})$ and $\mathcal{O}(r^{2d}\text{polylog}(r))$.

Remark 3.13. One of the advantages of our method over a vectorization-based approach is that the maps $\mathcal{A}$ and $\mathcal{A}_{2\text{nd}}$ require less memory to store than those considered in [34]. In particular, $\mathcal{A}$ requires $d'mn^\kappa$ entries to be stored in order to sketch a dimension of $m^{d'}$ and $\mathcal{A}_{2\text{nd}}$ requires $d'mn^\kappa + m^{d'}m_{2\text{nd}}$ in order to sketch a dimension of $m_{2\text{nd}}$. By contrast, the map considered in [34] requires $n^d m$ entries to reach a dimension of m. To make this more concrete, we consider the case where $d = 4$, $n = 40$, $\kappa = 2$, the final target dimension is $10,000$, and our maps are dense matrices with Gaussian entries. In this case, $\mathcal{A}_{2\text{nd}}$, with intermediate dimensions of $m_1 = m_2 = 250$ would require $40^2 \times 250 \times 2 + 250^2 \times 10,000 = 625,800,000$ random entries, with a storage cost of about 2.5 gigabytes assuming 32-bit floating point arithmetic and the SI meaning of the prefix giga as 10^9 bytes. The vectorization based approach requires $40^4 \times 10,000 = 25,600,000,000$ random entries at a storage cost of about 102.4 gigabytes. In Section 4.1, we will show that under these settings tensor recovery experiments using $\mathcal{A}_{2\text{nd}}$ have an identical recovery rate reliability in both the Gaussian case and SORS case when compared to those that use vectorization-based compression, despite the 40 times smaller memory requirement.

4. Low-rank tensor recovery

Low-rank tensor recovery is the task of recovering a low-rank (or approximately low-rank) tensor from a comparatively small number of possibly noisy linear measurements. This problem serves as a nice motivating example of where the use of modewise maps with the TRIP$(\delta, \mathbf{r})$ property can help alleviate the burdensome storage requirements of maps which require vectorization. Indeed, when the goal is compression, storing very large maps in memory as required by vectorization-based approaches is counterintuitive and often infeasible.

In the two-mode (matrix) case, the question of low-rank recovery from a small number of linear measurements is now well-known to be possible under various model assumptions on the measurements [11,10,35]. One of the standard approaches is so-called nuclear-norm minimization:

$$\hat{\mathcal{X}} = \underset{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \|\mathcal{X}\|_* \quad \text{subject to} \quad \mathcal{L}(\mathcal{X}) = y.$$

Since the nuclear norm is defined to be the sum of the singular values, it serves as a fairly good, computationally feasible proxy for rank. As in classical compressed sensing, an alternative to optimization-based reconstruction is the use of iterative solvers. One such approach is the Iterative Hard Thresholding (IHT) method [8,9,38] that finds a solution via the alternating updates

$$\begin{aligned} \mathcal{Y}^j &= \mathcal{X}^j + \mu_j \mathcal{L}^*(y - \mathcal{L}(\mathcal{X}^j)), \\ \mathcal{X}^{j+1} &= \mathcal{H}_{\mathbf{r}}\left(\mathcal{Y}^j\right), \end{aligned} \tag{22}$$

where $\mathcal{X}^0$ is initiated randomly. Here, $\mathcal{L}^*$ denotes the adjoint of the operator $\mathcal{L}$, and the function $\mathcal{H}_{\mathbf{r}}$ is a thresholding operator, which returns the closest rank $\mathbf{r}$ matrix via a truncated SVD. Results for IHT prove that sparse vector or low-rank matrix recovery is guaranteed when the measurement operator $\mathcal{L}$ satisfies various properties. For example, in the case of sparse vector recovery, the restricted isometry property is enough to guarantee accurate reconstruction [8]. In the low-rank matrix recovery case, measurements can be taken to be Gaussian [13], or satisfy various analogues of the restricted isometry property [38,7,41]. In what follows, for the sake of simplicity, we will focus on the case where $\mu_j = 1$, which is referred to as Classical IHT. However, our results can also be extended to Normalized TIHT where the step size μ_j takes a different value at each step. (See [34] and the references provided there.)

The iterative hard thresholding method has been extended to the tensor case ([18,34,19]). In this problem, one aims to recover an unknown tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ with e.g., HOSVD rank $\mathbf{r} = (r, \ldots, r)$, where $r \ll \min n_i$, from linear measurements of the form $\mathbf{y} = \mathcal{L}(\mathcal{X}) + \mathbf{e}$, where $\mathcal{L}$ is a linear map from $\mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{C}^m$, with $m \ll \prod_i n_i$, and $\mathbf{e}$ is an arbitrary noise vector. The iteration update is given by the same updates as

(22). The primary difference with the matrix case is in the thresholding operator $\mathcal{H}_{\mathbf{r}}$ that approximately computes the best rank $\mathbf{r}$ approximation of a given tensor. Unfortunately, exactly computing the best rank $\mathbf{r}$ approximation of a general tensor is NP-hard. However, it is possible to construct an operator $\mathcal{H}_{\mathbf{r}}$ in a way such that

$$\|\mathcal{Z} - \mathcal{H}_{\mathbf{r}}(\mathcal{Z})\|_F \leqslant C\sqrt{d}\|\mathcal{Z} - \mathcal{Z}_{\text{BEST}}\|_F, \tag{23}$$

where $\mathcal{Z}_{\text{BEST}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is the true best rank $\mathbf{r}$ approximation of $\mathcal{Z} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. (For details, please see [34] and the references therein.) For the rest of this section, we will always assume that $\mathcal{H}_{\mathbf{r}}$ is constructed in a way to satisfy (23).

The following theorem is the main result of [34]. It guarantees accurate reconstruction of a tensor $\mathcal{X}$ via TIHT guarantee when the measurement operator satisfies the TRIP$(\delta, 3\mathbf{r})$ property for a sufficiently small δ. Unfortunately, the condition (24), required by this theorem, is a bit stronger than (23), which is guaranteed to hold. As noted in [34], getting rid of the condition (24) appears to be difficult if not impossible. That said, (23) is a worst-case estimate, and in our numerical experiments we observe $\mathcal{H}_{\mathbf{r}}$ typically returns much better estimates and the condition (24) does often hold, especially in early iterations of the algorithm.

Theorem 4.1 *([34], Theorem 1). Let $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, let $0 < a < 1$, and let $\mathcal{L} : \mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{R}^m$ satisfy TRIP$(\delta, 3\mathbf{r})$ with $\delta < a/4$ for some $a \in (0, 1)$. Assume that $\mathbf{y} = \mathcal{L}(\mathcal{X}) + \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^m$ is an arbitrary noise vector, and let $\mathcal{X}^j$ and $\mathcal{Y}^j$ be defined as in (22). Assume that*

$$\|\mathcal{Y}^j - \mathcal{X}^{j+1}\|_F \leqslant (1 + \xi_a)\|\mathcal{Y}^j - \mathcal{X}\|_F, \quad \text{where} \quad \xi_a = \frac{a^2}{17(1 + \sqrt{1 + \delta_{3\mathbf{r}}})\|\mathcal{L}\|_{2\to 2}}. \tag{24}$$

Then

$$\|\mathcal{X}^{j+1} - \mathcal{X}\|_F \leqslant a^j\|\mathcal{X}^0 - \mathcal{X}\|_F + \frac{b_a}{1-a}\|\mathbf{e}\|_2,$$

where $b_a = 2\sqrt{1+\delta} + \sqrt{4\xi_a + 2\xi_a^2}\|\mathcal{L}\|_{2\to 2}$.

Theorem 4.1 shows that low-rank tensor recovery is possible when the measurements satisfy the TRIP$(\delta, 3\mathbf{r})$ property. In [34], the authors also show it is possible to randomly construct maps which satisfy this property with high probability. Unfortunately, these maps require first vectorizing the input tensor into a n^d-dimensional vector and then multiplying by an $m \times n^d$ matrix. This greatly limits the practical use of such maps since this matrix requires more memory than the original tensor. Thus, our results here for modewise TRIP are especially important and applicable in the tensor recovery setting. The following corollary, which shows that we may choose $\mathcal{L} = \mathcal{A}$ or $\mathcal{A}_{2nd}$ (as in (9) or (17)), now follows immediately from combining Theorem 4.1 with Theorems 3.3 and 3.8.

Corollary 4.2. *Assume the operator $\mathcal{L}$, is defined in one of the following ways:*

(a) *$\mathcal{L} = vect \circ \mathcal{A} \circ \mathcal{R}$, where $\mathcal{A}$ is defined as per (9), vect is a vectorization operator, and the matrices A_i satisfy the RIP$(\varepsilon, \mathcal{S}_{1,2})$, and $\delta = 4d'(3r)^d\varepsilon < a/4$.*
(b) *$\mathcal{L} = \mathcal{A}_{2nd}$ defined as in (17), its component matrices A_i satisfy RIP$(\varepsilon, \mathcal{S}_{1,2})$ property, $\delta = 12{d'}^2(3r)^d\varepsilon$, and A_{final} satisfies the RIP$(\delta/3, \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}})$ property.*

Consider the recovery problem from the noisy measurements $\mathbf{y} = \mathcal{L}(\mathcal{X}) + \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^m$ is an arbitrary noise vector. Let $0 < a < 1$, and let $\mathcal{X}^j$, and $\mathcal{Y}^j$ be defined as in (22), and assume that (24) holds. Then,

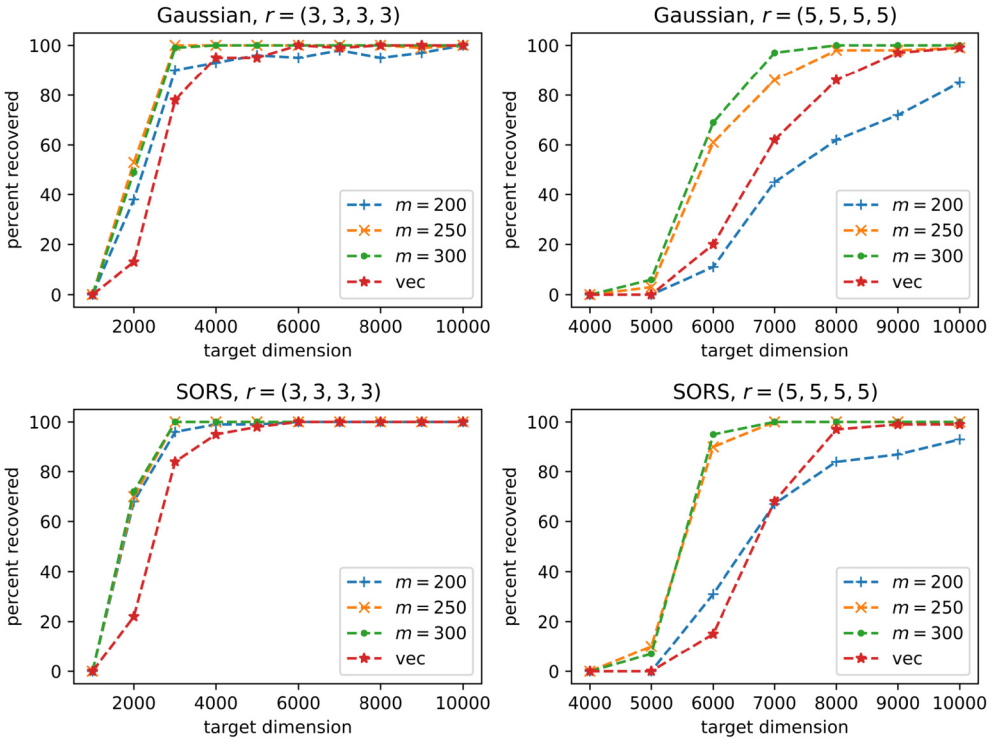

Fig. 1. Fraction of successfully recovered random tensors out of a random sample of 100 tensors in $\mathbb{R}^{40\times 40\times 40\times 40}$ with various intermediate dimensions. A run is considered successful if relative error is below 1% in at most 1000 iterations.

$$\|\mathcal{X}^{j+1} - \mathcal{X}\|_F \leqslant a^j \|\mathcal{X}^0 - \mathcal{X}\|_F + \frac{b_a}{1-a}\|\mathbf{e}\|,$$

where $b_a = 2\sqrt{1+\delta} + \sqrt{4\xi_a + 2\xi_a^2}\|\mathcal{A}\|_{2\to 2}$.

4.1. Experiments

In this section, we show that TIHT can be used with modewise measurement maps as defined in (22) can be used to successfully reconstruct low-rank tensors. In our experiments we will consider random four-mode tensors in $\mathbb{R}^{n\times n\times n\times n}$ for $n = 40$ and $n = 96$. In the case where $n = 40$, we will utilize both modewise Gaussian and SORS measurements. In the case where $n = 96$ we will only consider SORS measurements.[1]

In our experiments, we compare our two-step modewise approach to a vectorization based method. We consider the percentage of successfully recovered tensors from batches of 100 randomly generated low-rank tensors, as well as the average number of iterations used for recovery on the successful runs. In the case where $n = 40$, we consider the algorithm to have successfully recovered the tensor if the relative error falls below 1% in at most 1000 iterations. Similarly, in the case where $n = 96$, we consider the algorithm to have successfully recovered the tensor if the relative error falls below 5% in at most 1000 iterations. Compression in terms of final size of measurements over total number of entries in the true tensor ranges from about 0.04% to 0.4% depending on the choice for final sketching dimension. In all instances, we initialize the algorithm using a randomly generated low-rank tensor. In our experiments, apply the map from (17) with $\kappa = 2$. That is, we reshape a four-mode tensor whose modes are all of length n into a $n^2 \times n^2$ matrix, perform modewise measurements reducing each of the two reshaped modes to m, the choice for intermediate dimension, and then in the second stage, vectorize that result and compress it further to $m_{2\text{nd}}$, the final target dimension. In

[1] Our code is available at https://github.com/MichaelPerlmutter/ModewiseTrip.

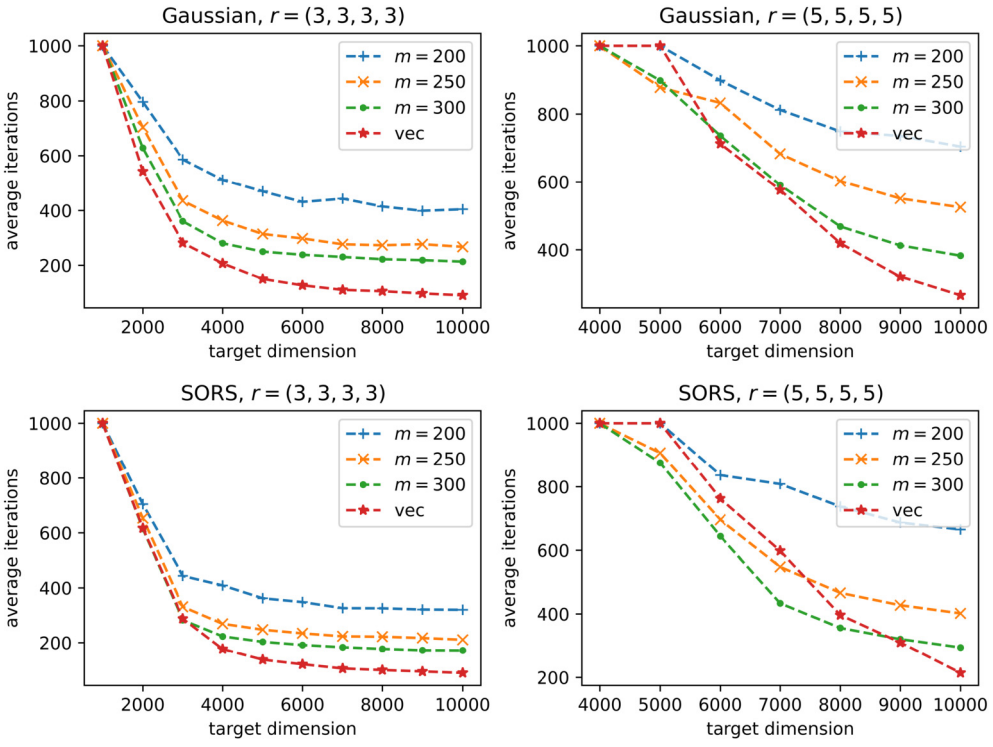

Fig. 2. Average number of iterations until convergence among the successful runs out of a random sample of 100 tensors in $\mathbb{R}^{40\times40\times40\times40}$ with various intermediate dimensions. A run is considered successful if relative error is below 1% in at most 1000 iterations.

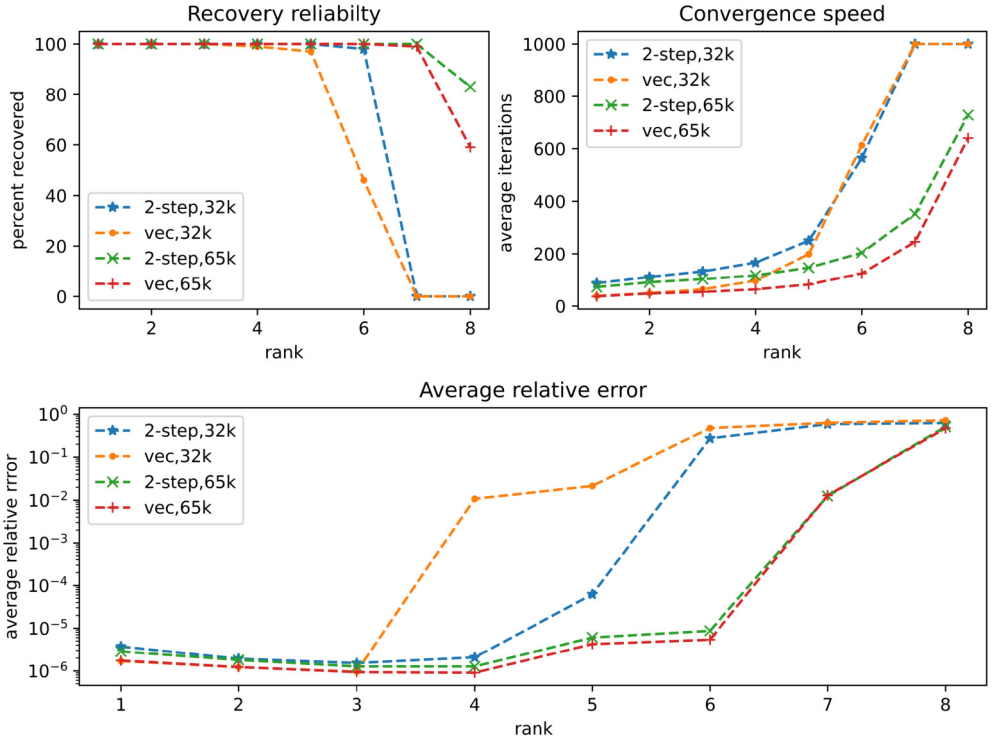

Fig. 3. Top row shows recovery reliability and convergence speed for trials consisting of 100 samples of random tensors in $\mathbb{R}^{96\times96\times96\times96}$ at various ranks. A run is successful if relative error reaches below 5% within 1000 iterations. Final target dimensions are $m = 32,768$ and $65,636$, for two-step method intermediate dimensions are 2048 and 4096 respectively. Bottom row is average relative error over 100 samples after 500 iterations.

our experiments, we consider a variety of intermediate dimensions to demonstrate the stability of advantage of the modewise measurements over the vectorized ones. In our experiments, for fair comparison, we will set the final embedding of our two-step process $m_{2\text{nd}}$ equal to the final embedding dimension of the vectorized method m_{final}. For a unified presentation, we will refer to the final dimension as m_0 in either case.

As noted in Remark 3.13, our proposed two-step method offers significant storage savings as compared to the vectorized based approach. For example, when we use Gaussian measurements with $m_0 = 10,000$ and $n = 40$, the vectorized computations had to be carried out using four NVIDIA v100 GPUs in parallel, whereas the two-step method easily fits on memory of one GPU. In the case where $n = 96$, generating, storing and applying Gaussian random matrices is impractical. For instance in the scenario we consider in Fig. 3, where $n = 96$ and $m_0 = 65,536$, we would need more than 22.25 terabytes to store the Gaussian map required for the vectorized approach. A two-step approach would require about 4.4 terabytes to store the measurement matrices.

We also note that we may apply SORS measurements to larger tensors than Gaussian measurements. This is because a fast Fourier transform enables SORs measurement matrices to be applied to the tensor without explicitly forming all the measurement maps. In particular, we need only store the sign changes which form the diagonal of matrix D, and store the choice of which rows are sampled from F to form matrix H (see Definition 3.5). This same size of problem requires about 67 megabytes storage in two-step method and 340 megabytes for the vectorized approach. Thus we restricted ourselves to SORS measurements for in the $n = 96$ case for both the vectorized and two-step approach.

As shown in Figs. 1 and 2 the two-step approach, when compared to the vectorized approach with the same choice of final sketching dimension, shows reliable recovery rate, and comparable convergence speed with both Gaussian and SORS measurements. Indeed, for some choices of intermediate and final sketching dimensions, modewise measurements empirically recover low-rank tensors more reliably than vectorized measurements (see Fig. 1, bottom row). We show that these advantages do not result in the need for a substantially increased number of iterations in order to achieve our convergence criteria. Across all considered scenarios, the average number of iterations to meet the convergence criteria can be bounded by two to eight times the number needed in the corresponding vectorized approach, depending on the choice m for intermediate sketching dimension. Interestingly, for some ranges of the intermediate and final target dimension m_0, and in the case of SORS measurements in the rank $(5, 5, 5, 5)$ instance, fewer iterations are needed (see Fig. 2, bottom right). Thus, modewise measurements are an effective, memory-efficient method of dimension reduction, and the choice of intermediate sketching dimension allows us to further balance trade-offs in terms of convergence speed and overall memory requirements, given a particular size and rank.

In Fig. 3, we investigate the performance of the algorithms as rank of the tensor increases. A larger tensor, with $n = 96$ enables us to consider a wider range of $\mathbf{r}$'s that are reasonably considered to be low-rank. We maintain the vectorized to two-step comparison, and also consider two different final sketching dimensions, $m_0 = 32,768$ and $65,536$. Due to the larger problem size and ranks, we scaled the convergence criteria to be 5% relative error in at most 1000 iterations for the comparison of recovery reliability and convergence speed. We observed in terms of performance of the algorithms a phase change as rank of the tensor increases for a fixed final sketching dimension. In particular, for $m_0 = 32,768$ the performance of vectorized and two-step approach empirically degrade significantly at $r = 6$ and 7 respectively for this convergence criteria. When doubling the sketching dimension to $m_0 = 65,536$ we see empirically that phase change that the drop in performance occurs at $r = 8$. The bottom row of Fig. 3 shows the shift in terms of average relative error for a fixed number of iterations. For the larger ranks and smaller sketching dimensions we observe that stagnation at non-global optima appears more likely and the runtimes required for acceptable recovery can become impractical. Both vectorized and two-step approaches have this feature, however for a fixed final sketching dimension, for some ranks near the threshold the two-step method performs incrementally better.

5. Proofs and theoretical guarantees

In this section, we will state auxiliary results that link the RIP property on a set $\mathcal{S}$ with the covering number of $\mathcal{S}$, and establish the covering number estimates for the subsets of interest.

5.1. Auxiliary results: RIP estimates

For a set $\mathcal{S}$ we let $\mathcal{N}(\mathcal{S}, t)$ denote its covering number, i.e., the minimal cardinality of a net contained in S, such that every element of S is within distance t of an element of the net. For further discussion of the covering numbers, please see [40]. The following proposition shows the estimates on the covering number of $\mathcal{S}$ can be used to show that maps constructed from subgaussian matrices have the RIP$(\varepsilon, \mathcal{S})$ property. Its proof, which is a generalization of the proof of [34, Theorem 2], can be found in Appendix C.

Proposition 5.1. *Suppose $A \in \mathbb{R}^{m \times n^\kappa}$ has i.i.d. $\frac{1}{m}$-subgaussian entries. Let $\mathcal{S} \subseteq \mathbb{R}^{n \times \cdots \times n}$ be a subset of unit norm κ-mode tensors and let $\mathcal{N}(\mathcal{S}, t)$ denote the covering number of $\mathcal{S}$. Then for any $0 < \eta, \varepsilon < 1$ and*

$$m \geqslant \tilde{C}\varepsilon^{-2} \max\left\{ \left(\int_0^1 \sqrt{\ln \mathcal{N}(\mathcal{S}, t)} dt \right)^2, 1, \ln(\eta^{-1}) \right\}, \tag{25}$$

for some suitably chosen constant $\tilde{C} > 0$, with probability at least $1 - \eta$, the map $\mathcal{A}(\mathcal{X}) = A(\mathrm{vect}(\mathcal{X}))$ has the RIP$(\varepsilon, \mathcal{S})$ property, i.e.,

$$(1 - \varepsilon)\|\mathcal{X}\|^2 \leqslant \|A(\mathrm{vect}(\mathcal{X}))\|^2 \leqslant (1 + \varepsilon)\|\mathcal{X}\|^2 \quad \text{for all } \mathcal{X} \in \mathcal{S}.$$

We also need an analogue of Proposition 5.1 that holds for SORS matrices. Such results are known in the literature, however all of them have additional logarithmic terms compared to the i.i.d. subgaussian case. We shall use the following result from [22] which is a refinement of Theorem 3.3 of [32].

Theorem 5.2 *(Theorem 9 of [22]). Suppose $A \in \mathbb{R}^{m \times n^\kappa}$ is a SORS matrix as per Definition 3.5 with $\Delta \leqslant C'$ for an absolute constant C'. Let $\mathcal{S}$ be a subset of $\mathbb{R}^{n^\kappa}$, and let ω denote the Gaussian width (see, e.g. [40]) of the projection of $\mathcal{S}$ onto the unit ball, $\widetilde{\mathcal{S}} := \left\{ \mathbf{x}/\|\mathbf{x}\|_2 \;\middle|\; \mathbf{x} \in \mathcal{S} \setminus \{\mathbf{0}\} \right\}$, i.e. $\omega := \sup_{\mathbf{x} \in \widetilde{\mathcal{S}}} \langle \mathbf{x}, \mathbf{g} \rangle_2$ where $\mathbf{g}$ is a standard Gaussian random vector. Let $0 < \eta, \varepsilon < 1$ and assume*

$$m \geqslant \tilde{C}\varepsilon^{-2}\omega^2 \ln^2\left[c_1 \omega^2 \ln(2\eta^{-1})\varepsilon^{-2}\right] \ln(2\eta^{-1}) \ln(2en^\kappa \eta^{-1}), \tag{26}$$

for some suitably chosen constants $\tilde{C}, c_1 > 0$. Then, with probability at least $1 - \eta$, the matrix A has the RIP$(\varepsilon, \mathcal{S})$ property, i.e.,

$$(1 - \varepsilon)\|\mathbf{x}\|^2 \leqslant \|A\mathbf{x}\|^2 \leqslant (1 + \varepsilon)\|\mathbf{x}\|^2 \quad \text{holds for all } \mathbf{x} \in \mathcal{S}.$$

Remark 5.3. It is known (see, e.g., Theorem 8.1.10 of [40]) that the Gaussian width $\omega(\mathcal{S})$ can be estimated by the same Dudley-type integral as used in (25). Namely, for any set $\mathcal{S}$,

$$\omega(\mathcal{S}) \leq \int_0^\infty \sqrt{\ln \mathcal{N}(\mathcal{S}, t)}\, dt.$$

Additionally, if $\mathcal{S}$ is a subset of the unit ball, we have $\ln(\mathcal{N}(\mathcal{S}, t)) = 0$ for $t > 2$, and therefore,

$$\omega(\mathcal{S}) \leq \int_0^2 \sqrt{\ln \mathcal{N}(\mathcal{S}, t)}\, dt.$$

5.2. Auxiliary results: covering estimates

The proofs of Corollaries 3.4 and 3.9 rely on applying Proposition 5.1 to the sets $\mathcal{S}_{1,2}$ and $\mathcal{B}_{\mathbf{r},R,\theta,\mu}$ defined in Definitions 3.1 and 3.7. The proofs of Corollaries 3.6 and 3.11 analogously follow from an application of Theorem 5.2. The following two lemmas provide covering estimates for these sets. Their proofs can be found in Appendix B.

Lemma 5.4 *(Covering number for very low rank tensors). The covering number for the set $\mathcal{S}_{1,2}$ defined in Definition 3.1 satisfies*

$$\mathcal{N}(\mathcal{S}_{1,2}, t) \leq \left(\left(\frac{6\kappa}{t}\right)^{\kappa n} + 1\right)^2.$$

Lemma 5.5. *For all $\theta \geqslant 0$, $0 < \varepsilon, \mu < 1$, $R \geqslant 1$, and all $\mathbf{r} = (r, r, \ldots, r) \in \mathbb{R}^d$, the set $\mathcal{B}_{R,\mu,\theta,\mathbf{r}} \subset \mathbb{R}^n$ defined in Definition 3.7 admits a covering with*

$$\mathcal{N}(\mathcal{B}_{R,\mu,\theta,\mathbf{r}}, \|\cdot\|_F, \varepsilon) \leqslant \left(\frac{6(d+1)}{\varepsilon}\right)^{r^d + rnd} \left(R^2 + \mu r\right)^{r^d d/2} \left(R^2 + \mu r^d\right)^{dnr/2} R^{(d-1)dnr}. \tag{27}$$

(Note that the right-hand side is independent of θ.)

Remark 5.6. If we set $\mu = \theta = 0$ and $R = 1$ we may obtain the covering number bound

$$\mathcal{N}(\mathcal{B}_{1,0,0,\mathbf{r}}, \|\cdot\|_F, \varepsilon) \leqslant \left(\frac{3(d+1)}{\varepsilon}\right)^{r^d + dnr},$$

via a trivial modification of the proof of Lemma 5.5 which doesn't require an application of Lemma B.1 when $\theta = 0$. This is the same as the estimate obtained in [34, Lemma 5] for $\mathcal{B}_{1,0,0,\mathbf{r}}$.

Remark 5.7. One may note that the right hand side of (27) increases rapidly as R grows large. Indeed, if we do not assume that orthogonality holds in (4), it is difficult if not impossible to obtain bounds independent of R. For instance, our result has similar R dependency to Lemma 2.6 of [19] which provides a covering number bound of tensors with bounded factors and low CP rank. Nevertheless, when we apply Lemma 5.4, we will set $R = (1 + \varepsilon)$ and so this dependence will not have a large influence on our main results such as Theorem 3.8.

5.3. Proof of Theorem 3.3 and Corollaries 3.4 and 3.6

In order to prove Theorem 3.3, it will be useful to write $\mathcal{A}$ as a composition of maps

$$\mathcal{A}(\mathcal{Y}) = \mathcal{A}_{d'}(\ldots(\mathcal{A}_1(\mathcal{Y}))), \text{ where } \mathcal{A}_i(\mathcal{Y}) = \mathcal{Y} \times_i A_i \text{ for } 1 \leqslant i \leqslant d'. \tag{28}$$

Our argument will be based on showing that $\mathcal{A}_i$ approximately preserves the norm of $\mathcal{A}_{i-1}(\ldots(\mathcal{A}_1(\mathring{\mathcal{X}})))$ for all $1 \leqslant i \leqslant d'$. We first note that by (8), we may still write $\mathring{\mathcal{X}}$ as a sum of r^d orthogonal tensors. This motivates Lemma 5.8 which shows that if a linear operator L on an inner product space V satisfies certain assumptions, then it approximately preserves the norm of orthogonal sums (up to a factor depending on the

number of terms). Lemma 5.9 then provides sufficient conditions for the assumptions of Lemma 5.8 to hold. Lastly, Lemma 5.10 will show that the image of the first $i-1$ compressions, $\mathcal{A}_{i-1}(\dots(\mathcal{A}_1(\mathring{\mathcal{X}})))$, satisfies these conditions and therefore that we may proceed inductively. The proofs of Lemmas 5.8, 5.9, and 5.10 are deferred to Appendix A.

Lemma 5.8. *Let $\mathcal{V}$ be an inner product space and let $\mathcal{L}$ be a linear operator on $\mathcal{V}$. Let $\mathcal{U} \subset \mathcal{V}$ be a subspace of $\mathcal{V}$ spanned by an orthonormal system $\{\mathbf{v}_1, \dots, \mathbf{v}_K\} \in \mathcal{V}$. Suppose that*

$$(1-\varepsilon)\|\mathbf{v}_i\|^2 \leqslant \|\mathcal{L}\mathbf{v}_i\|^2 \leqslant (1+\varepsilon)\|\mathbf{v}_i\|^2 \quad \text{for all } 1 \leqslant i \leqslant K, \tag{29}$$

and also that

$$(1-\varepsilon)\|\mathbf{v}_i \pm \mathbf{v}_j\|^2 \leqslant \|\mathcal{L}(\mathbf{v}_i \pm \mathbf{v}_j)\|^2 \leqslant (1+\varepsilon)\|\mathbf{v}_i \pm \mathbf{v}_j\|^2 \quad \text{for all } 1 \leqslant i, j \leqslant K. \tag{30}$$

Then we have

$$(1-K\varepsilon)\|\mathbf{w}\|^2 \leqslant \|\mathcal{L}\mathbf{w}\|^2 \leqslant (1+K\varepsilon)\|\mathbf{w}\|^2 \text{ for all } \mathbf{w} \in \mathcal{U}.$$

The next lemma checks that, if A_{i_0} satisfies RIP$(\varepsilon, \mathcal{S}_{1,2})$ property for some $1 \leqslant i_0 \leqslant d'$, then the operator $\mathcal{A}_{i_0}$ satisfies the conditions of Lemma 5.8 for the system of rank one component tensors that are produced by our reshaping procedure.

Lemma 5.9. *Let $\{\mathcal{V}_1, \dots, \mathcal{V}_K\} \in \mathbb{R}^{n \times \dots \times n}$ be an orthonormal system of rank one tensors of the form $\mathcal{V}_k = \bigcirc_{i=1}^{d'} \mathbf{v}_k^i$ where $\|\mathbf{v}_k^i\| = 1$ for all $1 \leqslant i \leqslant d'$. Let $1 \leqslant i_0 \leqslant d'$, suppose A_{i_0} has the RIP$(\varepsilon/2, \mathcal{S}_{1,2})$ property and assume that each $\mathbf{v}_k^{i_0}$ is an element of the set S_1 defined in Definition 3.1. Then the conditions (29) and (30) are satisfied for (the vectorizations of) these $\{\mathcal{V}_i\}_{i=1}^K$ and $\mathcal{L} = \mathcal{A}_{i_0}$ defined via $\mathcal{A}_{i_0}(\mathcal{X}) = \mathcal{X} \times_{i_0} A_{i_0}$.*

The next auxiliary lemma gives a formula for the tensor $\mathcal{Y}_t$ obtained by applying the first t of the maps $\mathcal{A}_i$. In particular, it shows that $\mathcal{Y}_t$ can be written as an orthogonal linear combination of $r^{\kappa(d'-t)}$ rank-one tensors of unit norm. Moreover, for each of the terms in this sum, the $(t+1)$-st component vector is $\mathring{\mathbf{u}}_{j_{t+1}}^{t+1}$ as defined in (8) and therefore is an element of the set S_1.

Lemma 5.10. *Let $\mathcal{Y}_0 = \mathring{\mathcal{X}}$ and $\mathcal{Y}_t := \mathcal{A}_t(\mathcal{Y}_{t-1}) = \mathcal{Y}_{t-1} \times_t A_t$ for all $t = 1, \dots, d'$. Then, for each $1 \leqslant t \leqslant d'-1$, we may write*

$$\mathcal{Y}_t = \sum_{j_{d'}=1}^{r^\kappa} \cdots \sum_{j_{t+1}=1}^{r^\kappa} \mathcal{C}_t(j_{t+1}, \dots, j_{d'}) \left[\left(\bigcirc_{i=1}^{t} \mathbf{v}^i_{j_{t+1},\dots,j_{d'}} \right) \bigcirc \left(\bigcirc_{i=t+1}^{d'} \mathring{\mathbf{u}}^i_{j_i} \right) \right], \tag{31}$$

where $\|\mathbf{v}^i_{j_{t+1},\dots,j_{d'}}\| = 1$ for all valid index subsets. (We note that the terms $\mathbf{v}^i_{j_{t+1},\dots,j_{d'}}$ implicitly depend on t. However, we suppress this dependence in order to avoid cumbersome notation.)

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. First, note that we can write $\mathcal{Y}_0 = \mathring{\mathcal{X}}$ as an orthogonal linear combination of r^d norm one terms of the form

$$\bigcirc_{i=1}^{d'} \mathring{\mathbf{u}}^i_{j_i}, \quad 1 \leqslant j_i \leqslant r^\kappa,$$

where each of the vectors $\mathring{\mathbf{u}}_{j_i}^i$, $1 \leqslant j_i \leqslant r^\kappa$, are obtained as the vectorization of a rank-one κ-mode tensor. Therefore, since A_1 satisfies RIP$(\varepsilon, \mathcal{S}_{1,2})$, Lemma 5.9 allows us to apply Lemma 5.8 to see

$$\|\mathcal{A}_1(\mathring{\mathcal{X}})\| \leqslant (1 + 2r^d\varepsilon)\|\mathring{\mathcal{X}}\|. \tag{32}$$

Next, we apply Lemma 5.10 and note that there are $r^{\kappa(d'-t)}$ terms appearing in the sum in (31). Therefore, Lemmas 5.8 and 5.9 allow us to see that

$$\|\mathcal{Y}_{t+1}\| \leqslant \left(1 + 2r^{\kappa(d'-t)}\varepsilon\right) \|\mathcal{Y}_t\| \tag{33}$$

for $1 \leqslant t \leqslant d' - 1$. Since $\mathcal{Y}_{d'} = \mathcal{A}(\mathring{\mathcal{X}})$, combining (32) and (33) implies that the operator $\mathcal{A}$ defined in (9) satisfies

$$\|\mathcal{A}(\mathring{\mathcal{X}})\| \leqslant \prod_{t=0}^{d'-1} \left(1 + 2r^{\kappa(d'-t)}\varepsilon\right) \|\mathring{\mathcal{X}}\|.$$

To complete the upper bound set $\alpha := 2r^d\varepsilon$ and note that $2\alpha < 1$. Then, since $r \geqslant 2$

$$\begin{aligned}
\prod_{t=0}^{d'-1} \left(1 + 2r^{\kappa(d'-t)}\varepsilon\right) &= \prod_{t=0}^{d'-1} (1 + \alpha r^{-t\kappa}) \\
&= 1 + \alpha \sum_{t=0}^{d'-1} r^{-t\kappa} + \alpha^2 \sum_{\substack{t_1,t_2=0: \\ t_1<t_2}}^{d'-1} r^{-(t_1+t_2)\kappa} + \ldots + \alpha^{d'} r^{-(1+\ldots+(d'-1))\kappa} \\
&\leq 1 + \alpha \sum_{t=0}^{d'-1} r^{-t\kappa} + \left(\alpha \sum_{t=0}^{d'-1} r^{-t\kappa}\right)^2 + \ldots + \left(\alpha \sum_{t=0}^{d'-1} r^{-t\kappa}\right)^{d'} \\
&\leq 1 + \alpha \sum_{t=0}^{\infty} 2^{-t} + \left(\alpha \sum_{t=0}^{\infty} 2^{-t}\right)^2 + \ldots + \left(\alpha \sum_{t=0}^{\infty} 2^{-t}\right)^{d'} \\
&\leq 1 + 2\alpha + (2\alpha)^2 + \ldots + (2\alpha)^{d'} \\
&\leq 1 + 2d'\alpha \\
&= 1 + 4d'r^d\varepsilon
\end{aligned}$$

which completes the proof of the upper bound. The proof of the lower bound is nearly identical. □

We will now prove Corollaries 3.4 and 3.6.

Proof of Corollary 3.4. We first note that Lemma 5.4 implies that the integral from (25) can be bounded as

$$\int_0^1 \sqrt{\ln \mathcal{N}(\mathcal{S}_{1,2}, t)}\, dt \leq C \int_0^1 \sqrt{\kappa n \ln(6\kappa/t)}\, dt \leq C\sqrt{\kappa n \ln(\kappa)}. \tag{34}$$

Since the set $\mathcal{S}_{1,2}$ contains (reshaped) unit norm κ-tensors, the assumption that m satisfies (13) implies that each of the A_i will satisfy the assumptions of Proposition 5.1 with η/d' in place of η and $\varepsilon^{-2} = (d/\kappa)^2 r^{2d}\delta^{-2}$. Therefore, by the union bound, we have that all of the A_i will satisfy RIP$(\varepsilon, \mathcal{S}_{1,2})$ with probability at least $1 - \eta$, and so the result now follows from Theorem 3.3. □

Proof of Corollary 3.6. To estimate the Gaussian width $\mathcal{S}_{1,2}$, we use (34), Lemma 5.4, and Remark 5.3 to see that

$$\omega(\mathcal{S}_{1,2}) \leq \int_0^2 \sqrt{\ln \mathcal{N}(\mathcal{S}_{1,2},t)}\, dt \leq \int_0^1 \sqrt{\ln \mathcal{N}(\mathcal{S}_{1,2},t)}\, dt + \sqrt{\ln \mathcal{N}(\mathcal{S}_{1,2},1)}$$
$$\leq C''\sqrt{\kappa n \ln(\kappa)} + C'''\sqrt{\kappa n \ln(\kappa)} = C\sqrt{\kappa n \ln(\kappa)}.$$

We now observe that the assumption (15) allows us to apply Theorem 5.2 with η/d' in place of η and $\varepsilon^{-2} = (d/\kappa)^2 r^{2d}\delta^{-2}$. Therefore, analogous to the proof of Corollary 3.4, we conclude the proof of by taking the union bound and applying Theorem 3.3. □

5.4. Proof of Theorem 3.8 and Corollaries 3.9 and 3.11

The key to proving Theorem 3.8 is Lemma 5.11, which shows that the output of the first compression step $\mathcal{A}(\mathcal{R}(\mathcal{X}))$ lies in a set of nearly orthogonal tensors introduced in Definition 3.7, and Lemma 5.5 which bounds the covering numbers for such tensors. We can then get TRIP by applying Proposition 5.1 to the vectorization of $\mathcal{A}(\mathcal{R}(\mathcal{X}))$.

Lemma 5.11. *Let $\mathcal{X}$ be a unit norm d-mode tensor with HOSVD rank at most $\mathbf{r} = (r,\dots,r)$. Let $\mathcal{A}$ be defined as in (9), assume that the matrices A_i have the RIP$(\varepsilon,\mathcal{S}_{1,2})$ property for all $1 \leqslant i \leqslant d'$, and that $\delta = 12d'r^d\varepsilon < 1$. Then, the d'-mode tensor $\mathcal{A}(\mathring{\mathcal{X}}) \in \mathbb{R}^{m\times\dots\times m}$ is an element of the set $\mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}$ (as per Definition 3.7). Additionally, let $\tilde{\mathcal{B}}_{\varepsilon,\delta,r} := \left\{ \frac{\mathcal{X}}{\|\mathcal{X}\|_{\mathrm{F}}} \;\middle|\; \mathcal{X} \in \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'} \right\}$ and suppose that $m \geqslant r^{d-1}$. Then,*

$$\mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, \|\cdot\|_{\mathrm{F}}, t\right) \leqslant \left(\frac{9(d/\kappa+1)}{t}\right)^{r^d + r^\kappa md/\kappa} \left((1+\varepsilon)^2 + \varepsilon r^d\right)^{dmr^\kappa/\kappa} (1+\varepsilon)^{d^2 m r^\kappa \kappa^{-2}}$$

holds for all $t > 0$, and $1 > \delta > 0$.

Proof. As in (8) we set $\mathring{\mathcal{X}} = \mathcal{R}(\mathcal{X})$, and write

$$\mathring{\mathcal{X}} = \sum_{j_{d'}=1}^{r^\kappa} \dots \sum_{j_1=1}^{r^\kappa} \mathring{C}(j_1,\dots,j_{d'}) \bigcirc_{\ell=1}^{d'} \mathring{\mathbf{u}}_{j_\ell}^{\ell},$$

where, for all i, the vectors $\{\mathring{\mathbf{u}}_{j_i}^i\}_{j_i=1}^{r^\kappa}$ are mutually orthogonal. Recall that by (10), we have

$$\mathcal{A}(\mathring{\mathcal{X}}) = \sum_{j_{d'}=1}^{r^\kappa} \dots \sum_{j_1=1}^{r^\kappa} \mathring{C}(j_1,\dots,j_{d'}) \left(A_1\mathring{\mathbf{u}}_{j_1}^1 \bigcirc \dots \bigcirc A_{d'}\mathring{\mathbf{u}}_{j_{d'}}^{d'}\right).$$

By RIP$(\varepsilon,\mathcal{S}_{1,2})$ property, $\|A_i\mathring{\mathbf{u}}_{j_i}^i\| \leqslant (1+\varepsilon)$ for all $1 \leqslant i \leqslant d'$ and all $1 \leqslant j_i \leqslant r^\kappa$. Additionally, by Lemma A.1 (stated in Appendix A) we have

$$|\langle A_i\mathring{\mathbf{u}}_{j_i}^i, A_i\mathring{\mathbf{u}}_{j_i'}^i\rangle| < \varepsilon$$

for all $1 \leqslant i \leqslant d'$ and all $1 \leqslant j_i, j_i' \leqslant r^\kappa$ such that $j_i' \neq j_i$. The properties of the core tensor (c) and (d) are preserved under the action of $\mathcal{A}$ and are satisfied for a unit norm, tensor in the HOSVD standard form. Finally, Theorem 3.3 implies that $\mathcal{A}$ satisfies the TRIP$(\delta/3,\mathbf{r})$ property which in turn guarantees that $\mathcal{A}(\mathring{\mathcal{X}})$ will also satisfy property (e) of Definition 3.7 with $\theta = 1-\delta/3$.

Applying Lemma 5.5 (with $R = (1+\varepsilon)$ and m, d', r^κ, and t in place of n, d, r, and ε) and the assumption $m \geqslant r^{d-1}$ we obtain

$$\mathcal{N}\left(\mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}, \|\cdot\|_{\mathrm{F}}, t\right) \leqslant \left(\frac{6(d/\kappa+1)}{t}\right)^{r^d + r^\kappa m d/\kappa} \left((1+\varepsilon)^2 + \varepsilon r^d\right)^{dmr^\kappa/\kappa} (1+\varepsilon)^{d^2 m r^\kappa \kappa^{-2}}.$$

Furthermore, a geometric rescaling argument implies that

$$\mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, \|\cdot\|_{\mathrm{F}}, t\right) = \mathcal{N}\left(\left\{\frac{\mathcal{X}}{\|\mathcal{X}\|_{\mathrm{F}}} \;\middle|\; \mathcal{X} \in \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}\right\}, \|\cdot\|_{\mathrm{F}}, t\right) \leqslant \mathcal{N}\left(\mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}, \|\cdot\|_{\mathrm{F}}, 2t/3\right)$$

hols for all $\delta \leqslant 1$. The stated result now follows. □

Theorem 3.8 now follows from a direct application of Lemma 5.11 and Theorem 3.3.

Proof of Theorem 3.8. Theorem 3.3 implies that $\mathcal{A}$ satisfies the TRIP$(\delta/3, \mathbf{r})$ property and Lemma 5.11 implies that $\mathcal{A}(\mathring{\mathcal{X}})$ belongs to the set $\mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}$. Therefore, we have

$$\|\mathcal{A}_{2nd}(\mathcal{X})\|^2 \leqslant (1+\delta/3)\|\mathcal{A}(\mathring{\mathcal{X}})\| \leqslant (1+\delta/3)^2\|\mathcal{X}\|^2 \leqslant (1+\delta)\|\mathcal{X}\|^2,$$

and

$$(1-\delta)\|\mathcal{X}\|^2 \leqslant (1-\delta/3)^2\|\mathcal{X}\|^2 \leqslant (1-\delta/3)\|\mathcal{A}(\mathring{\mathcal{X}})\| \leqslant \|\mathcal{A}_{2nd}(\mathcal{X})\|^2. \quad \square$$

The proofs of Corollaries 3.9 and 3.11 require an additional estimate bounding the Gaussian width of $\tilde{\mathcal{B}}_{\varepsilon,\delta,r}$ from Lemma 5.11.

Remark 5.12. Let $\tilde{\mathcal{B}}_{\varepsilon,\delta,r} \subset \mathbb{R}^{m\times\cdots\times m}$ be the set of d'-mode tensors defined as per Lemma 5.11 and Definition 3.7. We can see that

$$\int_0^1 \sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, t\right)}\, dt \leqslant \int_0^2 \sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, t\right)}\, dt \leqslant \int_0^1 \sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, t\right)}\, dt + \sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, 1\right)}.$$

If $m \geqslant r^{d-1}$, then we may apply Lemma 5.11 and use the concavity of $\sqrt{\cdot}$ to see that

$$\begin{aligned}\int_0^1 \sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, t\right)}\, dt &\leqslant \int_0^1 \sqrt{\left(r^d + \frac{r^\kappa m d}{\kappa}\right)\ln\left(\frac{9(d/\kappa+1)}{t}\right)}\, dt \\ &\quad + \int_0^1 \sqrt{\frac{dmr^\kappa}{\kappa}\ln\left((1+\varepsilon)^2 + \varepsilon r^d\right)}\, dt + \int_0^1 \sqrt{\frac{d^2 m r^\kappa}{\kappa^2}\ln(1+\varepsilon)}\, dt \\ &\leq C\sqrt{r^d + \frac{dmr^\kappa}{\kappa}}\sqrt{\ln\left(\frac{d}{\kappa}+1\right)} + \sqrt{\frac{dmr^\kappa}{\kappa}\ln((1+\varepsilon)^2 + \varepsilon r^d)} + \sqrt{\frac{d^2 m r^\kappa}{\kappa^2}\ln(1+\varepsilon)}.\end{aligned}$$

Furthermore, we also have

$$\begin{aligned}\sqrt{\ln \mathcal{N}\left(\tilde{\mathcal{B}}_{\varepsilon,\delta,r}, 1\right)} &\leqslant C'\sqrt{r^d + \frac{dmr^\kappa}{\kappa}}\sqrt{\ln\left(\frac{d}{\kappa}+1\right)} \\ &\quad + \sqrt{\frac{dmr^\kappa}{\kappa}\ln((1+\varepsilon)^2 + \varepsilon r^d)} + \sqrt{\frac{d^2 m r^\kappa}{\kappa^2}\ln(1+\varepsilon)}.\end{aligned}$$

Proof of Corollary 3.9. Similar to the proof of Corollary 3.4, the assumption that m satisfies (18), implies that with probability at least $1-\eta/2$, all of the A_i satisfy the RIP$(\varepsilon, \mathcal{S}_{1,2})$ property, $\delta = 12d'r^d\varepsilon < 1$. The assumption (18) also implies $m \geqslant r^{d-1}$. Therefore, by Proposition 5.1, the estimate for the Dudley-type integral from Remark 5.12, and the linearity of $A_{2\text{nd}} \circ \text{vect}$ we see that $A_{2\text{nd}} \circ \text{vect}$ satisfies the RIP$\big(\delta/3, \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}\big)$ property with probability at least $1-\eta/2$ as long as

$$m_{2\text{nd}} \geqslant C\delta^{-2} \max\left\{\left(r^d + \frac{dmr^\kappa}{\kappa}\right)\ln\left(\frac{d}{\kappa}+1\right) + \frac{dmr^\kappa}{\kappa}\ln\left(1+\delta r^d\right) + \frac{d^2mr^\kappa\delta}{\kappa^2}, \ln\left(\frac{2}{\eta}\right)\right\}.$$

The result now follows by applying Theorem 3.8. □

Proof of Corollary 3.11. Repeating the arguments used in the proof of Corollary 3.6, we see that the assumption that m satisfies (20), implies that with probability at least $1-\eta/2$, all of the A_i satisfy the RIP$(\varepsilon, \mathcal{S}_{1,2})$ property with $\delta = 12d'r^d\varepsilon < 1$. We also note that (20) implies that $m \geqslant r^{d-1}$ so that we may apply Remark 5.12. Thus, (21) and Theorem 5.2 imply that $A_{2\text{nd}}$ satisfies the RIP$\big(\delta/3, \mathcal{B}_{1+\varepsilon,\varepsilon,1-\delta/3,\mathbf{r}'}\big)$ property with probability at least $1-\eta/2$. Therefore, the result now follows from Theorem 3.8. □

6. Conclusion and future work

In this paper, we have proved that several modewise linear maps (with subgaussian and subsampled from the orthogonal ensemble – e.g., discrete Fourier – measurements) have the TRIP for tensors with low-rank HOSVD decompositions. Our measurements maps require significantly less memory than previous works such as [34] and [19] that establish TRIP for vectorized measurements. We also note that unlike other closely related works such as [23] and [21] that establish modewise Johnson-Lindenstrauss embeddings, our results hold for all low-HOSVD rank tensors whereas previous work focuses on finite sets or for tensors lying in a low-dimensional vector space. In our experiments, we have demonstrated that we are able to recover low-rank tensors from a compressed representation produced via two-step modewise measurements. Moreover, we show that we are able to achieve such recovery from a lower compressed dimension than with purely vectorized measurements, establishing yet another advantage.

A natural direction for future work would involve extending these results to other tensor formats including, e.g., tensors which admit compact tensor train, (hierarchical) Tucker, and/or CP decompositions instead. Additional projects of value might include parallel implementations of the TIHT algorithm using modewise maps that fully leverage their structure, as well as more memory efficient TIHT variants which reconstruct the factors of a given low-rank tensor from its measurements instead of reconstructing the entire tensor in uncompressed form. Indeed, such a memory efficient TIHT implementation in combination with using modewise measurements would allow for memory efficient low-rank tensor reconstruction from the measurement stage all the way through reconstruction of the final approximation in compressed form.

Appendix A. The proof of Lemmas 5.8, 5.9, and 5.10

The proof of Lemma 5.8 requires the following well-known auxiliary lemma. For completeness, we provide a short proof below.

Lemma A.1. *Let $\mathcal{V}$ be an inner product space and let $\mathcal{L}$ be a linear operator on $\mathcal{V}$. Let $\{\mathbf{v}_1, \dots, \mathbf{v}_K\}$ be a finite orthonormal system in $\mathcal{V}$ (that is, $\|\mathbf{v}_i\| = 1$ for all i and $\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0$ for all $i \neq j$). Suppose that*

$$(1-\varepsilon)\|\mathbf{v}_i \pm \mathbf{v}_j\|^2 \leqslant \|\mathcal{L}(\mathbf{v}_i \pm \mathbf{v}_j)\|^2 \leqslant (1+\varepsilon)\|\mathbf{v}_i \pm \mathbf{v}_j\|^2 \quad \text{for all } 1 \leqslant i, j \leqslant K, i \neq j.$$

Then

$$|\langle \mathcal{L}\mathbf{v}_i, \mathcal{L}\mathbf{v}_j\rangle| \leqslant \varepsilon \quad \textit{for all } i \neq j.$$

Proof. Let $i \neq j$. Then,

$$\begin{aligned}
4\langle \mathcal{L}\mathbf{v}_i, \mathcal{L}\mathbf{v}_j\rangle &= \|\mathcal{L}(\mathbf{v}_i + \mathbf{v}_j)\|^2 - \|\mathcal{L}(\mathbf{v}_i - \mathbf{v}_j)\|^2 \\
&\leqslant (1+\varepsilon)\|\mathbf{v}_i + \mathbf{v}_j\|^2 - (1-\varepsilon)\|\mathbf{v}_i - \mathbf{v}_j\|^2 \\
&= (1+\varepsilon)(\|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 + 2\langle \mathbf{v}_i, \mathbf{v}_j\rangle) - (1-\varepsilon)(\|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 - 2\langle \mathbf{v}_i, \mathbf{v}_j\rangle) \\
&= 4\langle \mathbf{v}_i, \mathbf{v}_j\rangle + 2\varepsilon(\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2) \\
&= 4\varepsilon,
\end{aligned}$$

where the last inequality follows from the fact that $\|\mathbf{v}_1\|^2 = \|\mathbf{v}_2\|^2 = 1$ and $\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0$. Thus, $\langle \mathcal{L}\mathbf{v}_i, \mathcal{L}\mathbf{v}_j\rangle \leqslant \varepsilon$. The reverse inequality is similar. □

We may now prove Lemma 5.8.

The Proof of Lemma 5.8. We argue by induction on K. When $K = 1$, the result is immediate from (29) and the fact that $\mathcal{L}$ is linear. Now assume the result is true for $K-1$. An arbitrary element of $\mathcal{U}$ may be written as

$$\mathbf{w} = \sum_{i=1}^{K} c_i \mathbf{v}_i$$

where $c_1, \ldots, c_K$ are scalars. We will write $\mathbf{w} = \mathbf{w}_{K-1} + c_K \mathbf{v}_K$, where

$$\mathbf{w}_{K-1} := \sum_{i=1}^{K-1} c_i \mathbf{v}_i.$$

By construction, we have

$$\|\mathcal{L}\mathbf{w}\|^2 = \|\mathcal{L}\mathbf{w}_{K-1}\|^2 + \|c_K \mathcal{L}\mathbf{v}_K\|^2 + 2c_K\langle \mathcal{L}\mathbf{w}_{K-1}, \mathcal{L}\mathbf{v}_K\rangle.$$

We may use the inequality $2ab \leqslant a^2 + b^2$ along with Lemma A.1 to see

$$\begin{aligned}
2c_K\langle \mathcal{L}\mathbf{w}_{K-1}, \mathcal{L}\mathbf{v}_K\rangle = \sum_{i=1}^{K-1} 2c_i c_K \langle \mathcal{L}\mathbf{v}_i, \mathcal{L}\mathbf{v}_K\rangle &\leqslant \varepsilon \sum_{i=1}^{K-1} 2c_K c_i \\
&\leqslant \varepsilon \sum_{i=1}^{K-1} (c_K^2 + c_i^2) \leqslant (K-1)\varepsilon c_K^2 + \varepsilon \sum_{i=1}^{K-1} c_i^2.
\end{aligned} \tag{35}$$

By the inductive assumption,

$$\|\mathcal{L}\mathbf{w}_{K-1}\|^2 \leqslant (1 + (K-1)\varepsilon)\|\mathbf{w}_{K-1}\|^2 = (1 + (K-1)\varepsilon) \sum_{i=1}^{K-1} c_i^2.$$

Thus,

$$\begin{aligned}\|\mathcal{L}\mathbf{w}\|^2 &\leqslant (1+(K-1)\varepsilon)\sum_{i=1}^{K-1} c_i^2 + (1+\varepsilon)c_K^2 + (K-1)\varepsilon c_K^2 + \varepsilon\sum_{i=1}^{K-1} c_i^2 \\ &= (1+K\varepsilon)\sum_{i=1}^{K} c_i^2 = (1+K\varepsilon)\|\mathbf{w}\|^2.\end{aligned} \tag{36}$$

The reverse inequality is similar. □

Proof of Lemma 5.9. Without loss of generality, we consider the case where $i_0 = 1$. By assumption, we have $\mathbf{v}_k^1 \in S_1$ for all $1 \leqslant k \leqslant K$. Therefore, since A_1 is assumed to have the RIP$(\varepsilon/2, \mathcal{S}_{1,2})$ property, we have

$$1-\varepsilon/2 = (1-\varepsilon/2)\|\mathbf{v}_k^1\|^2 \leqslant \|A_1\mathbf{v}_k^1\|^2 \leqslant (1+\varepsilon/2)\|\mathbf{v}_k^1\|^2 = 1+\varepsilon/2.$$

Now, (29) follows from the fact that

$$\left\|\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_k^i\right)\right\| = \|(A_1\mathbf{v}_k^1)\bigcirc\mathbf{v}_k^2\bigcirc\ldots\bigcirc\mathbf{v}_k^{d'}\| = \|A_1\mathbf{v}_k^1\|\|\mathbf{v}_k^2\|\ldots\|\mathbf{v}_k^{d'}\| = \|A_1\mathbf{v}_k^1\|,$$

and the fact that the $\mathbf{v}_k^i$ have norm one. To prove (30), we let $1 \leqslant k_1, k_2 \leqslant K$ and recall that $\mathcal{V}_{k_1} = \bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i$ and $\mathcal{V}_{k_2} = \bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i$, where each of the $\mathbf{v}_{k_1}^i$ and $\mathbf{v}_{k_2}^i$ have norm one. If $k_1 = k_2$, (30) follows immediately from (29). Otherwise, we may use the assumption that $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ form an orthonormal system to see

$$0 = \langle\mathcal{V}_{k_1}, \mathcal{V}_{k_2}\rangle = \left\langle\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i, \bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\rangle = \prod_{i=1}^{d'}\langle\mathbf{v}_{k_1}^i, \mathbf{v}_{k_2}^i\rangle.$$

This implies that there exists an i such that $\langle\mathbf{v}_{k_1}^i, \mathbf{v}_{k_2}^i\rangle = 0$. If $\langle\mathbf{v}_{k_1}^1, \mathbf{v}_{k_2}^1\rangle = 0$, then since A_1 satisfies RIP$(\varepsilon, \mathcal{S}_{1,2})$, we may apply Lemma A.1, and the Cauchy-Schwarz inequality to see that

$$\begin{aligned}\left|\left\langle\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right), \mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right)\right\rangle\right| &= |\langle A_1\mathbf{v}_{k_1}^1, A_1\mathbf{v}_{k_2}^1\rangle||\langle\mathbf{v}_{k_1}^2, \mathbf{v}_{k_2}^2\rangle|\ldots|\langle\mathbf{v}_{k_1}^{d'}, \mathbf{v}_{k_2}^{d'}\rangle| \\ &\leqslant (\varepsilon/2)\prod_{i=2}^{d'}\|\mathbf{v}_{k_1}^i\|\|\mathbf{v}_{k_2}^i\| \\ &= \varepsilon/2.\end{aligned}$$

On the other hand, if $\langle\mathbf{v}_{k_1}^\ell, \mathbf{v}_{k_2}^\ell\rangle = 0$ for some $\ell \neq 1$ then we have

$$\left|\left\langle\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right), \mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right)\right\rangle\right| = |\langle A_1\mathbf{v}_{k_2}^1, A_1\mathbf{v}_{k_2}^1\rangle||\langle\mathbf{v}_{k_1}^2, \mathbf{v}_{k_2}^2\rangle|\ldots|\langle\mathbf{v}_{k_1}^{d'}, \mathbf{v}_{k_2}^{d'}\rangle| = 0.$$

Therefore, in either case we have

$$\begin{aligned}\left\|\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i+\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right)\right\|^2 &= \left\|\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right)\right\|^2+\left\|\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right)\right\|^2+2\left\langle\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right),\mathcal{A}_1\left(\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right)\right\rangle\\ &\leqslant (1+\varepsilon/2)\left(\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right\|^2+\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2\right)+\varepsilon\\ &= (1+\varepsilon/2)\left(\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right\|^2+\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2\right)+(\varepsilon/2)\left(\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right\|^2+\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2\right)\\ &= (1+\varepsilon)\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i+\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2,\end{aligned}$$

where in the last equality, we used orthogonality to see

$$\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i+\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2=\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_1}^i\right\|^2+\left\|\bigcirc_{i=1}^{d'}\mathbf{v}_{k_2}^i\right\|^2.$$

The reverse inequality is similar. □

Proof of Lemma 5.10. We argue by induction. When $t=0$, the decomposition (31) follows immediately from (8) with $\mathcal{C}_0=\mathring{\mathcal{C}}$ and the fact that $\mathcal{Y}_0=\mathring{\mathcal{X}}$.

Now suppose that the result is true for t for some $0\leqslant t\leqslant d-2$. Then,

$$\begin{aligned}\mathcal{Y}_{t+1} &= \mathcal{Y}_t\times_{t+1}\mathcal{A}_{t+1}\\ &= \sum_{j_{d'}=1}^{r^\kappa}\cdots\sum_{j_{t+1}=1}^{r^\kappa}\mathcal{C}_t(j_{t+1},\dots,j_{d'})\left[\left(\bigcirc_{i=1}^{t}\mathbf{v}_{j_{t+1},\dots,j_{d'}}^i\right)\bigcirc\left(A_{t+1}\mathring{\mathbf{u}}_{j_{t+1}}^{t+1}\right)\bigcirc\left(\bigcirc_{i=t+2}^{d'}\mathring{\mathbf{u}}_{j_i}^i\right)\right].\end{aligned}$$

Therefore, summing over j_{t+1}-st mode and normalizing yields

$$\mathcal{Y}_{t+1}=\sum_{j_{d'}=1}^{r^\kappa}\cdots\sum_{j_{t+2}=1}^{r^\kappa}\mathcal{C}_{t+1}(j_{t+2},\dots,j_{d'})\left[\left(\bigcirc_{i=1}^{t+1}\mathbf{v}_{j_{t+2},\dots,j_{d'}}^i\right)\bigcirc\left(\bigcirc_{i=t+2}^{d'}\mathring{\mathbf{u}}_{j_i}^i\right)\right],$$

where the new terms $\mathbf{v}_{j_{t+2},\dots,j_{d'}}^i$ (with one less subscript) are defined as

$$\mathbf{v}_{j_{t+2},\dots,j_{d'}}^i=\frac{\tilde{\mathbf{v}}_{j_{t+2},\dots,j_{d'}}^i}{\left\|\tilde{\mathbf{v}}_{j_{t+2},\dots,j_{d'}}^i\right\|}\quad\text{and}\quad\mathcal{C}_{t+1}(j_{t+2},\dots,j_{d'})=\left\|\tilde{\mathbf{v}}_{j_{t+2},\dots,j_{d'}}^i\right\|,$$

where

$$\tilde{\mathbf{v}}_{j_{t+2},\dots,j_{d'}}^i=\sum_{j_{t+1}=1}^{r^\kappa}\mathcal{C}_t(j_{t+1},\dots,j_{d'})\left(\bigcirc_{i=1}^{t}\mathbf{v}_{j_{t+1},\dots,j_{d'}}^i\right)\bigcirc A_{t+1}\mathring{\mathbf{u}}_{j_{t+1}}^{t+1}.\quad\square$$

Appendix B. The proof of Lemmas 5.4 and 5.5

The proof of Lemma 5.4. Classical results [40] utilizing volumetric estimates show that

$$\mathcal{N}(\mathbb{S}^{n-1},\|\cdot\|_2,t)\leqslant\left(\frac{3}{t}\right)^n.\tag{37}$$

Let $\mathcal{N}_1$ be the k-fold Kronecker product of $t/2\kappa$-nets for $\mathbb{S}^{n-1}$. By (37), we may choose $\mathcal{N}_1$ to have cardinality at most $(6\kappa/t)^{\kappa n}$. Moreover, for any $\mathcal{X} = \bigotimes_{i=1}^{\kappa} u^i \in S_1$ (defined by (11)),

$$\begin{aligned}
\inf_{\tilde{\mathcal{X}}\in\mathcal{N}_1} \|\tilde{\mathcal{X}} - \mathcal{X}\|_{\mathrm{F}} &\leq \inf_{\substack{\tilde{\mathcal{X}}\in\mathcal{N}_1 \\ \tilde{\mathcal{X}}=\otimes_{i=1}^{\kappa}\tilde{u}^i}} \left\| \bigotimes_{i=1}^{\kappa} u^i - \bigotimes_{i=1}^{\kappa} \tilde{u}^i \right\|_{\mathrm{F}} \\
&\leq \inf_{\substack{\tilde{\mathcal{X}}\in\mathcal{N}_1 \\ \tilde{\mathcal{X}}=\otimes_{i=1}^{\kappa}\tilde{u}^i}} \sum_{j=1}^{\kappa} \left\| \bigotimes_{i=1}^{j-1} u^i \bigotimes_{i=j}^{\kappa} \tilde{u}^i - \bigotimes_{i=1}^{j} u^i \bigotimes_{i=j+1}^{\kappa} \tilde{u}^i \right\|_{\mathrm{F}} \\
&= \inf_{\substack{\tilde{\mathcal{X}}\in\mathcal{N}_1 \\ \tilde{\mathcal{X}}=\otimes_{i=1}^{\kappa}\tilde{u}^i}} \sum_{j=1}^{\kappa} \left(\prod_{i=1}^{j-1} \|\tilde{u}^i\|_2 \right) \|\tilde{u}^j - u^j\|_2 \left(\prod_{i=j+1}^{\kappa} \|u^i\|_2 \right) \\
&\leq \kappa \frac{t}{2\kappa}.
\end{aligned}$$

Now, let $\mathcal{S}'_{1,2}$ be the set of nonzero $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2$ such that each $\mathcal{X}_i \in S_1 \cup \{\mathbf{0}\}$ and has $\langle \mathcal{X}_1, \mathcal{X}_2 \rangle = 0$. For a given $\mathcal{X} \in \mathcal{S}'_{1,2}$, we may set $\tilde{X} = \tilde{X}_1 + \tilde{X}_2$, where for $i = 1, 2$ $\tilde{X}_i$ is the best $(\mathcal{N}_1 \cup \{0\})$-approximation of $\mathcal{X}_i$, and note that

$$\|\mathcal{X} - \tilde{\mathcal{X}}\|_F \leq \|\mathcal{X}_1 - \tilde{\mathcal{X}}_1\|_F + \|\mathcal{X}_2 - \tilde{\mathcal{X}}_2\|_F \leq t.$$

Thus, $\mathcal{N}_2 := (\mathcal{N}_1 \cup \{\mathbf{0}\}) + (\mathcal{N}_1 + \cup\{\mathbf{0}\}) = \{\mathcal{X}_1 + \mathcal{X}_2 \mid X_1, X_2 \in \mathcal{N}_1 + \cup\{\mathbf{0}\}\} \setminus \{\mathbf{0}\}$ is a t-net of $\mathcal{S}'_{1,2}$ with cardinality at most $\left(\left(\frac{6\kappa}{t}\right)^{\kappa n} + 1\right)^2$. Lastly, we note that each element of $\mathcal{S}'_{1,2}$ has norm at least one and that $\mathcal{S}_{1,2}$ is the projection of $\mathcal{S}'_{1,2}$ onto the unit sphere. Therefore, the projection of $\mathcal{N}_2$ onto the unit sphere is t-net for $\mathcal{S}_{1,2}$. □

The following technical lemma will be used in the proof of Lemma 5.5.

Lemma B.1. *Let $A \subseteq B \subseteq \mathbb{R}^n$, and suppose that $C \subseteq B$ is an $\varepsilon/2$-net of B. Then, there exists an ε-net $C' \subseteq A$ of A with cardinality $|C'| \leqslant |C|$.*

Proof. We will construct C' from C as follows. First, let $\tilde{C}$ be the subset of C whose elements are all at least $\varepsilon/2$ away from A,

$$\tilde{C} := \left\{ \mathbf{x} \;\middle|\; \mathbf{x} \in C \text{ and } \inf_{\mathbf{y}\in A} \|\mathbf{x} - \mathbf{y}\|_2 \geqslant \varepsilon/2 \right\}.$$

Next, for each $\mathbf{x} \in C \setminus \tilde{C}$ let $\mathbf{x}' \in A$ be any point of A satisfying $\|\mathbf{x} - \mathbf{x}'\|_2 < \varepsilon/2$, and then set

$$C' := \bigcup_{\mathbf{x}\in C\setminus\tilde{C}} \mathbf{x}' \subseteq A.$$

Note that $|C'| \leqslant |C|$ by construction.

To see that C' is an ε-net of A, choose any $\mathbf{y} \in A \subseteq B$ and let $\mathbf{x} \in C$ be a point satisfying $\|\mathbf{y} - \mathbf{x}\| < \varepsilon/2$. Noting that $\mathbf{x} \notin \tilde{C}$, we can see that there is a $\mathbf{x}' \in C'$ such that $\|\mathbf{x} - \mathbf{x}'\|_2 < \varepsilon/2$. Therefore, by the triangle inequality,

$$\|\mathbf{y} - \mathbf{x}'\|_2 \leqslant \|\mathbf{y} - \mathbf{x}\|_2 + \|\mathbf{x} - \mathbf{x}'\|_2 < \varepsilon.$$

This establishes the desired result. □

The proof of Lemma 5.5. Our argument is based on the proof of [34, Lemma 5], with necessary modifications to account for the fact that the tensor factors are not orthogonal.

Part 1: Construction of the net. An arbitrary element of $\mathcal{B}_{R,\mu,\theta,\mathbf{r}}$ can be written as $\mathcal{X} = \mathcal{C} \times_1 V^1 \times_2 \ldots \times_d V^d$ where the core tensor $\mathcal{C} \in \mathbb{R}^{r \times \cdots \times r}$ is a d-mode tensor with Frobenius norm one and the factor matrices $V^i \in \mathbb{R}^{n \times r}$ have columns $\mathbf{v}_j^i$ with norm at most R. The set of all core tensors satisfying the orthogonality condition (d) is isometric to a subset of the unit ball in $\mathbb{R}^{r^d}$. Therefore, the admissible core tensors admit an ε_1-net $\mathcal{N}_1$ of the cardinality at most $(3/\varepsilon_1)^{r^d}$ by [34, Lemma 1]. We define the $\|\cdot\|_{1,2}$ norm by $\|V\|_{1,2} := \max_j \|\mathbf{v}_j\|_2$ and note that by construction we have $\|V^i\|_{1,2} \leqslant R$. Therefore, our admissible factor matrices satisfying condition (b) have an ε_2-net (with respect to the $\|\cdot\|_{1,2}$ norm) of the cardinality at most $(3R/\varepsilon_2)^{nr}$ again by [34, Lemma 1]. We now define

$$\mathcal{N} := \left\{ \overline{\mathcal{C}} \times_1 \overline{V}^1 \ldots \times_d \overline{V}^d : \overline{\mathcal{C}} \in \mathcal{N}_1 \text{ and } \overline{V}^i \in \mathcal{N}_2 \right\}, \quad |\mathcal{N}| \leq \left(\frac{3}{\varepsilon_1}\right)^{r^d} \left(\frac{3R}{\varepsilon_2}\right)^{dnr}.$$

Going forward we will prove that $\mathcal{N}$ above is an $\varepsilon/2$-net of $\mathcal{B}_{R,\mu,0,\mathbf{r}}$ for suitable choices of ε_1 and ε_2. The result will then follow from noting $\mathcal{B}_{R,\mu,\theta,\mathbf{r}} \subseteq \mathcal{B}_{R,\mu,0,\mathbf{r}}$ and applying Lemma B.1.

Part 2: Term by term approximation. Let us take an arbitrary element of $\mathcal{X} \in \mathcal{B}_{R,\mu,0,\mathbf{r}}$ and consider its component-wise approximation in $\overline{\mathcal{X}} \in \mathcal{N}$:

$$\mathcal{X} = \mathcal{C} \times_1 V^1 \times_2 \ldots \times_d V^d \in \mathcal{B}_{R,\mu,0,\mathbf{r}} \quad \text{and} \quad \overline{\mathcal{X}} = \overline{\mathcal{C}} \times_1 \overline{V}^1 \times_2 \ldots \times_d \overline{V}^d \in \mathcal{N},$$

where $\|\overline{\mathcal{C}} - \mathcal{C}\|_F \leqslant \varepsilon_1$ and $\|\overline{V}^i - V^i\|_{1,2} \leqslant \varepsilon_2$ for all $1 \leqslant i \leqslant d$. The triangle inequality implies that

$$\|\overline{\mathcal{X}} - \mathcal{X}\|_F \leqslant \sum_{j=0}^{d} \|\mathcal{T}_j\|_F \tag{38}$$

where $\mathcal{T}_0 = (\overline{\mathcal{C}} - \mathcal{C}) \times_1 \overline{V}^1 \times_2 \ldots \times_d \overline{V}^d$ and for $1 \leqslant j \leqslant d$,

$$\mathcal{T}_j = \mathcal{C} \times_1 V^1 \ldots \times_{j-1} V^{j-1} \times_j W^j \times_{j+1} \overline{V}^{j+1} \ldots \times_d \overline{V}^d \quad \text{for } 1 \leqslant j \leqslant d,$$

where $W^j := V^j - \overline{V}^j$.

Part 3: Bounding $\mathcal{T}_j$ for $\mathbf{j = 1, \ldots, d}$. For $1 \leqslant j \leqslant d$, we can expand

$$\begin{aligned} \|\mathcal{T}_j\|_F^2 &= \sum_{t_d=1}^{n} \cdots \sum_{t_1=1}^{n} |\mathcal{T}_j(t_1, \ldots, t_d)|^2 \\ &= \sum_{\substack{t_i=1 \\ 1 \leqslant i \leqslant d}}^{n} \left| \sum_{k_d=1}^{r} \cdots \sum_{k_1=1}^{r} \mathcal{C}(k_1, \ldots, k_d) \left[\prod_{i=1}^{j-1} v_{k_i}^i(t_i) \right] w_{k_j}^j(t_j) \left[\prod_{i=j+1}^{d} \overline{v}_{k_i}^i(t_i) \right] \right|^2, \end{aligned} \tag{39}$$

where $\mathbf{v}_1^i, \ldots, \mathbf{v}_r^i$ denote the columns of V^i, and similarly $\mathbf{w}_1^i, \ldots, \mathbf{w}_r^i$ and $\overline{\mathbf{v}}_1^i, \ldots, \overline{\mathbf{v}}_r^i$ denote the columns of W^i and $\overline{V}^i$. Exchanging the sums, we can rewrite (39) in the following way:

$$\sum_{\substack{l_i=1\\1\leqslant i\leqslant d}}^{r}\sum_{\substack{k_i=1\\1\leqslant i\leqslant d}}^{r}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)\sum_{\substack{t_i=1\\1\leqslant i\leqslant d}}^{n}\left[\prod_{i=1}^{j-1}v^i_{l_i}(t_i)v^i_{k_i}(t_i)\right]w^j_{l_j}(t_j)w^j_{k_j}(t_j)\left[\prod_{i=j+1}^{d}\overline{v}^i_{k_i}(t_i)\overline{v}^i_{k_i}(t_i)\right]$$

$$=\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d}}^{r}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)\left[\prod_{i=1}^{j-1}\langle\mathbf{v}^i_{k_i},\mathbf{v}^i_{\ell_i}\rangle\right]\langle\mathbf{w}^j_{k_j},\mathbf{w}^j_{\ell_j}\rangle\left[\prod_{i=j+1}^{d}\langle\overline{\mathbf{v}}^i_{k_i},\overline{\mathbf{v}}^i_{\ell_i}\rangle\right].$$

We estimate scalar products by $|\langle\mathbf{w}^i_{k_i},\mathbf{w}^i_{\ell_i}\rangle|\leq\|\mathbf{w}^i_{k_i}\|\|\mathbf{w}^i_{\ell_i}\|\leq\varepsilon_2^2$ and

$$|\langle\overline{\mathbf{v}}^i_{k_i},\overline{\mathbf{v}}^i_{\ell_i}\rangle|,|\langle\mathbf{v}^i_{k_i},\mathbf{v}^i_{\ell_i}\rangle|\leqslant\begin{cases}R^2, & \text{if } k_i=l_i,\\ \mu, & \text{otherwise}\end{cases}. \tag{40}$$

Therefore, for any $1\leq j\leq d$, we have

$$\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d}}^{r}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)\left[\prod_{i=1}^{j-1}\langle\mathbf{v}^i_{k_i},\mathbf{v}^i_{\ell_i}\rangle\right]\langle\mathbf{w}^j_{k_j},\mathbf{w}^j_{\ell_j}\rangle\left[\prod_{i=j+1}^{d}\langle\overline{\mathbf{v}}^i_{k_i},\overline{\mathbf{v}}^i_{\ell_i}\rangle\right]$$

$$=\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d\\k_i=\ell_i\forall i\neq j}}^{r}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)\left[\prod_{i=1}^{j-1}\langle\mathbf{v}^i_{k_i},\mathbf{v}^i_{\ell_i}\rangle\right]\langle\mathbf{w}^j_{k_j},\mathbf{w}^j_{\ell_j}\rangle\left[\prod_{i=j+1}^{d}\langle\overline{\mathbf{v}}^i_{k_i},\overline{\mathbf{v}}^i_{\ell_i}\rangle\right]$$

$$+\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d\\\exists i\neq j\text{ s.t. }k_i\neq\ell_i}}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)\left[\prod_{i=1}^{j-1}\langle\mathbf{v}^i_{k_i},\mathbf{v}^i_{\ell_i}\rangle\right]\langle\mathbf{w}^j_{k_j},\mathbf{w}^j_{\ell_j}\rangle\left[\prod_{i=j+1}^{d}\langle\overline{\mathbf{v}}^i_{k_i},\overline{\mathbf{v}}^i_{\ell_i}\rangle\right]$$

$$\leqslant R^{2(d-1)}\varepsilon_2^2\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d\\k_i=\ell_i\forall i\neq j}}^{r}\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)+\mu R^{2(d-2)}\varepsilon_2^2\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d}}^{r}|\mathcal{C}(k_1,\dots,k_d)\mathcal{C}(l_1,\dots,l_d)|$$

since $\mu<1\leqslant R$ by assumption. Hence, we have

$$\begin{aligned}\|\mathcal{T}_j\|_F^2 \;\leqslant\; & R^{2(d-1)}\varepsilon_2^2\sum_{\substack{k_i=1\\1\leq i\leq d,i\neq j}}^{r}\sum_{k_j=1}^{r}\sum_{\ell_j=1}^{r}\mathcal{C}(k_1,\dots,k_j,\dots,k_d)\mathcal{C}(k_1,\dots,l_j,\dots,k_d)\\ &+\mu R^{2(d-2)}\varepsilon_2^2\sum_{\substack{l_i=1,k_i=1\\1\leqslant i\leqslant d}}^{r}|\mathcal{C}(\ell_1,\dots,\ell_d)\mathcal{C}(k_1,\dots,k_d)|\\ =\; & R^{2(d-1)}\varepsilon_2^2\sum_{k_i=1}^{d}|\mathcal{C}(k_1,\dots,k_j,\dots,k_d)|^2+\mu R^{2(d-2)}\varepsilon_2^2\left(\sum_{\substack{\ell_i=1\\1\leq i\leq d}}^{r}|\mathcal{C}(\ell_1,\dots,\ell_d)|\right)^2,\end{aligned}$$

where in the last step we have used the fact that by the orthogonality property (d)

$$\sum_{\substack{k_i=1\\i\neq j}}^{r}\mathcal{C}(k_1,\dots,k_j,\dots,k_d)\mathcal{C}(k_1,\dots,l_j,\dots,k_d)=0\quad\text{unless } k_j=\ell_j,$$

to see that

$$\sum_{\substack{k_i=1\\ i\neq j}}^{r}\sum_{k_j=1}^{r}\sum_{\ell_j=1}^{r}\mathcal{C}(k_1,\dots,k_j,\dots,k_d)\mathcal{C}(k_1,\dots,l_j,\dots,k_d)=\sum_{k_i=1}^{d}|\mathcal{C}(k_1,\dots,k_j,\dots,k_d)|^2.$$

Recalling that all of our core tensors are unit norm, and appealing to Cauchy-Schwarz now allows us to see that

$$\begin{aligned}\|\mathcal{T}_j\|_F^2 &\leqslant R^{2(d-1)}\varepsilon_2^2+\mu R^{2(d-2)}\varepsilon_2^2 r^d\|\mathcal{C}\|_F^2\\ &\leq \varepsilon_2^2 R^{2d-4}(R^2+\mu r^d).\end{aligned}\tag{41}$$

Part 4: Bounding $\mathcal{T}_0$. We note that for any $1\leqslant i\leqslant d$, the $\|\cdot\|_F\to\|\cdot\|_F$ operator norm of the operator $\mathcal{X}\to\mathcal{X}\times_i V^i$ is the same as the ℓ_2-operator norm of the matrix V^i acting on $\mathbb{R}^r$. Next, we observe that for all $\mathbf{x}=(x_1,\dots,x_r)\in\mathbb{R}^r$ with $\|\mathbf{x}\|_2=1$, we have

$$\|V^i\mathbf{x}\|_2^2=\sum_{k,l=1}^{r}\langle\mathbf{v}_k^i,\mathbf{v}_l^i\rangle x_l x_k=\sum_{k=1}^{r}\langle\mathbf{v}_k^i,\mathbf{v}_k^i\rangle x_k^2+\sum_{k\neq l=1}^{r}\langle\mathbf{v}_k^i,\mathbf{v}_l^i\rangle x_k x_l.$$

Thus, bounding the coefficients by (40) and using Cauchy–Schwarz we have that

$$\|V^i\mathbf{x}\|_2^2\leqslant R^2\sum_{k=1}^{r}x_k^2+\mu\sum_{k\neq l=1}^{r}|x_k||x_l|\leqslant R^2\|\mathbf{x}\|_2^2+\mu\left(\sum_{k=1}^{r}|x_k|\right)^2\leqslant(R^2+r\mu)\|\mathbf{x}\|_2^2.$$

Therefore, since $\|\mathcal{C}-\overline{\mathcal{C}}\|_F^2\leqslant\varepsilon_1$,

$$\|\mathcal{T}_0\|_F^2=\|(\mathcal{C}-\overline{\mathcal{C}})\times_1\overline{V}_1\times_2\dots\times_d\overline{V}_d\|_F^2\leqslant(R^2+r\mu)^d\varepsilon_1^2.$$

Part 5: Conclusion and cardinality estimate. By (38), we get that $\|\mathcal{X}-\overline{\mathcal{X}}\|_F\leq\varepsilon/2$ if each $\|\mathcal{T}_j\|\leq\varepsilon/2(d+1)$. That is, taking $(R^2+\mu r)^{d/2}\varepsilon_1:=\varepsilon/2(d+1)$ and $\varepsilon_2R^{d-2}(R^2+\mu r^d)^{1/2}:=\varepsilon/2(d+1)$, we get

$$|\mathcal{N}|\leq\left(\frac{3}{\varepsilon_1}\right)^{r^d}\left(\frac{3R}{\varepsilon_2}\right)^{dnr}=\left(\frac{6(d+1)}{\varepsilon}\right)^{r^d+rnd}\left((R^2+\mu r)^{d/2}\right)^{r^d}\left(R^{d-1}(R^2+\mu r^d)^{1/2}\right)^{dnr}.\tag{42}$$

This concludes the proof of Lemma 5.5. □

Appendix C. The proof of Proposition 5.1

Proposition 5.1 follows by essentially repeating the proof of Theorem 2 of [34]. We include the argument here for completeness. The key probabilistic component of the proof is the supremum of chaos inequality proved in [25]:

Theorem C.1 *([25], Theorem 3.1). Let $\mathcal{A}$ be a collection of matrices, which size is measured through the following three quantities E,V and U defined as*

$$\begin{aligned}E(\mathcal{A})&:=\gamma_2(\mathcal{A})(\gamma_2(\mathcal{A})+d_F(\mathcal{A}))+d_F(\mathcal{A})d_2(\mathcal{A})\\ V(\mathcal{A})&:=d_2(\mathcal{A})(\gamma_2(\mathcal{A})+d_F(\mathcal{A})),\quad \textit{and}\\ U(\mathcal{A})&:=d_2^2(\mathcal{A}),\end{aligned}\tag{43}$$

where

$$d_F(\mathcal{A}) := \sup_{A\in\mathcal{A}} \|A\|_F, \quad d_2(\mathcal{A}) := \sup_{A\in\mathcal{A}} \|A\|_{2\to 2}$$

and $\gamma_2(\mathcal{A}) = \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$ *is Talagrand's functional. Let* $\boldsymbol{\xi}$ *be an* L*-subgaussian random vector whose entries* ξ_j *are independent, mean-zero and variance-1. Then, for* $t > 0$*,*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} \left|\|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2\right| \geqslant c_1 E(\mathcal{A}) + t\right) \leqslant 2\exp\left(-c_2 \min\left\{\frac{t^2}{V(\mathcal{A})^2}, \frac{t}{U(\mathcal{A})}\right\}\right),$$

where the constants c_1, c_2 *depend only on the subgaussian constant* L*.*

The proof of Proposition 5.1. Observe that

$$\varepsilon_S := \sup_{\mathcal{X}\in S, \|\mathcal{X}\|=1} \left|\|A(\mathrm{vect}(\mathcal{X}))\|^2 - 1\right| = \sup_{\mathcal{X}\in S, \|\mathcal{X}\|=1} \left|\|V_{\mathcal{X}}\boldsymbol{\xi}\|^2 - \mathbb{E}\|V_{\mathcal{X}}\boldsymbol{\xi}\|^2\right|,$$

where $\xi \in \mathbb{R}^{n^\kappa m}$ is a vector with i.i.d. $\frac{1}{m}$-subgaussian entries, and $V_{\mathcal{X}} \in \mathbb{R}^{m\times n^\kappa m}$ is a block-diagonal matrix

$$V_{\mathcal{X}} = \frac{1}{\sqrt{m}} \begin{bmatrix} \mathrm{vect}(\mathcal{X})^T & 0 & \dots & 0 \\ 0 & \mathrm{vect}(\mathcal{X})^T & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathrm{vect}(\mathcal{X})^T \end{bmatrix}.$$

Let $\mathcal{V} = \mathcal{V}(\mathcal{S})$ be the set of all such matrices $V_{\mathcal{X}}$ where $\mathcal{X} \in \mathcal{S}$. We will now apply Theorem C.1. It is easy to check (see [34, Theorem 3]) that

$$d_2(\mathcal{V}) = m^{-1/2} \text{ and } d_F(\mathcal{V}) = 1, \tag{44}$$

therefore, by Theorem C.1,

$$\mathbb{P}\left\{\varepsilon_{\mathcal{S}} \geqslant c_1\left(\gamma_2^2 + \gamma_2 + \frac{1}{\sqrt{m}}\right) + t\right\} \leq 2\exp\left(-c_2 \min\left\{mt, \frac{\sqrt{m}t^2}{1+\gamma_2}\right\}\right) \text{ for all } t > 0, \tag{45}$$

where $\gamma_2 = \gamma_2(\mathcal{V})$.

For any set $\mathcal{S}_0$, the Talagrand functional $\gamma_2(\mathcal{S}_0)$ is a functional of $\mathcal{S}_0$ which can be bounded by the Dudley-type integral

$$\gamma_2(\mathcal{S}_0) \leq C \int_0^{d_2(\mathcal{S}_0)} \sqrt{\ln \mathcal{N}(\mathcal{S}_0, \|\cdot\|_{2\to 2}, t)}\, dt. \tag{46}$$

We need to consider $\mathcal{S}_0 = \mathcal{V}(\mathcal{S})$. We note that due to (44),

$$\mathcal{N}(\mathcal{V}(\mathcal{S}), \|\cdot\|_{2\to 2}, u) \leqslant \mathcal{N}(\mathcal{S}, \|\cdot\|_{\mathrm{F}}, \sqrt{m}u),$$

and so by a change of variables

$$\gamma_2(\mathcal{V}(\mathcal{S})) \leqslant C\frac{1}{\sqrt{m}} \int_0^1 \sqrt{\ln(\mathcal{N}(\mathcal{S}, \|\cdot\|_{\mathrm{F}}, u))}\, du.$$

This implies that $\gamma_2(\mathcal{V}) \leq C\tilde{C}^{-1/2}\varepsilon$ by the main condition on m given in the statement of Proposition 5.1. Choosing $\tilde{C}$ large enough, this ensures that $\gamma_2 \leq \varepsilon/6c_1$ for any $c_1 > 1$.

Taking $t = \varepsilon/2$ and using $\varepsilon < 1$, we can now rewrite (45) as

$$\mathbb{P}\left\{\varepsilon_{\mathcal{S}} \geqslant c_1\left(\left[\frac{\varepsilon}{6c_1}\right]^2 + \frac{\varepsilon}{6c_1} + \frac{1}{\sqrt{m}}\right) + \frac{\varepsilon}{2}\right\} \leq 2\exp\left(-C_2 m\varepsilon^2\right)$$

If $\tilde{C}$ is chosen large enough, then $m \geqslant \tilde{C}\varepsilon^{-2}\ln\eta^{-1}$ implies that the probability bound is at most η and $m \geqslant \tilde{C}\varepsilon^{-2}$ implies that $m^{-1/2} \leq \varepsilon/6c_1$. This concludes the proof of Proposition 5.1. □

References

[1] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, J. Comput. Syst. Sci. 66 (4) (2003) 671–687.
[2] T.D. Ahle, M. Kapralov, J.B. Knudsen, R. Pagh, A. Velingker, D.P. Woodruff, A. Zandieh, Oblivious sketching of high-degree polynomial kernels, in: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2020, pp. 141–160.
[3] S. Bamberger, F. Krahmer, R. Ward, Johnson-Lindenstrauss embeddings with Kronecker structure, arXiv preprint, arXiv: 2106.13349, 2021.
[4] R.G. Baraniuk, M.B. Wakin, Random projections of smooth manifolds, Found. Comput. Math. 9 (1) (Feb. 2009) 51–77.
[5] M.H. Beck, A. Jäckle, G.A. Worth, H.-D. Meyer, The multiconfiguration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets, Phys. Rep. 324 (1) (2000) 1–105.
[6] J.A. Bengua, H.N. Phien, H.D. Tuan, M.N. Do, Efficient tensor completion for color image and video recovery: low-rank tensor train, IEEE Trans. Image Process. 26 (5) (2017) 2466–2479.
[7] J.D. Blanchard, J. Tanner, K. Wei, Cgiht: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion, Inf. Inference 4 (4) (2015) 289–327.
[8] T. Blumensath, M.E. Davies, Iterative hard thresholding for compressed sensing, Appl. Comput. Harmon. Anal. 27 (3) (2009) 265–274.
[9] T. Blumensath, M.E. Davies, Normalized iterative hard thresholding: guaranteed stability and performance, IEEE J. Sel. Top. Signal Process. 4 (2) (2010) 298–309.
[10] E.J. Candes, Y. Plan, Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements, arXiv preprint, arXiv:1001.0339, 2010.
[11] E.J. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717–772.
[12] E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, in: A Journal Issued by the Courant Institute of Mathematical Sciences, Commun. Pure Appl. Math. 59 (8) (2006) 1207–1223.
[13] A. Carpentier, A.K. Kim, An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution, Stat. Sin. 28 (3) (2018) 1371–1393.
[14] S. Dasgupta, A. Gupta, An elementary proof of the Johnson-Lindenstrauss lemma, Int. Comput. Sci. Inst., Technical Report 22 (1) (1999) 1–5.
[15] S. Foucart, Hard thresholding pursuit: an algorithm for compressive sensing, SIAM J. Numer. Anal. 49 (6) (2011) 2543–2563.
[16] S. Foucart, Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants, in: Approximation Theory XIII: San Antonio 2010, Springer, 2012, pp. 65–77.
[17] S. Foucart, H. Rauhut, A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis, Springer, New York, New York, NY, 2013.
[18] J.H.d.M. Goulart, G. Favier, An iterative hard thresholding algorithm with improved convergence for low-rank tensor recovery, in: 2015 23rd European Signal Processing Conference (EUSIPCO), IEEE, 2015, pp. 1701–1705.
[19] R. Grotheer, S. Li, A. Ma, D. Needell, J. Qin, Iterative hard thresholding for low CP-rank tensor models, arXiv preprint, arXiv:1908.08479, 2019.
[20] R. Grotheer, A. Ma, D. Needell, S. Li, J. Qin, Stochastic iterative hard thresholding for low Tucker rank tensor recovery, in: Proc. Information Theory and Applications, 2020.
[21] M. Iwen, D. Needell, E. Rebrova, A. Zare, Lower memory oblivious (tensor) subspace embeddings with fewer random bits: modewise methods for least squares, SIAM J. Matrix Anal. Appl. 42 (2021) 376–416.
[22] M.A. Iwen, B. Schmidt, A. Tavakoli, On fast Johnson-Lindernstrauss embeddings of compact submanifolds of $\mathbb{R}^n$ with boundary, arXiv:2110.04193, 2021.
[23] R. Jin, T.G. Kolda, R. Ward, Faster Johnson-Lindenstrauss transforms via Kronecker products, Inf. Inference 10 (4) (2021) 1533–1562, https://doi.org/10.1093/imaiai/iaaa028.
[24] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (3) (2009) 455–500.
[25] F. Krahmer, S. Mendelson, H. Rauhut, Suprema of chaos processes and the restricted isometry property, Commun. Pure Appl. Math. 67 (11) (2014) 1877–1904.
[26] J. Liu, P. Musialski, P. Wonka, J. Ye, Tensor completion for estimating missing values in visual data, IEEE Trans. Pattern Anal. 35 (1) (2012) 208–220.

[27] C. Lubich, From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis, European Mathematical Society, 2008.
[28] O.A. Malik, S. Becker, Guarantees for the Kronecker fast Johnson–Lindenstrauss transform using a coherence and sampling argument, Linear Algebra Appl. 602 (2020) 120–137.
[29] J. Matoušek, On variants of the Johnson–Lindenstrauss lemma, Random Struct. Algorithms 33 (2) (2008) 142–156.
[30] Q. Mo, S. Li, New bounds on the restricted isometry constant δ_{2k}, Appl. Comput. Harmon. Anal. 31 (3) (2011) 460–468.
[31] D. Needell, J.A. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. 26 (3) (2009) 301–321.
[32] S. Oymak, B. Recht, M. Soltanolkotabi, Isometric sketching of any set via the restricted isometry property, Inf. Inference 7 (4) (2018) 707–726.
[33] B.T. Rakhshan, G. Rabusseau, Tensorized random projections, arXiv preprint, arXiv:2003.05101, 2020.
[34] H. Rauhut, R. Schneider, Ž. Stojanac, Low rank tensor recovery via iterative hard thresholding, Linear Algebra Appl. 523 (2017) 220–262.
[35] B. Recht, M. Fazel, P.A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev. 52 (3) (2010) 471–501.
[36] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, M. Pontil, Multilinear multitask learning, in: International Conference on Machine Learning, 2013, pp. 1444–1452.
[37] Y. Sun, Y. Guo, C. Luo, J. Tropp, M. Udell, Low-rank Tucker approximation of a tensor from streaming data, SIAM J. Math. Data Sci. 2 (4) (2020) 1123–1150.
[38] J. Tanner, K. Wei, Normalized iterative hard thresholding for matrix completion, SIAM J. Sci. Comput. 35 (5) (2013) S104–S125.
[39] M.A.O. Vasilescu, D. Terzopoulos, Multilinear Independent Components Analysis, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 547–553.
[40] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, vol. 47, Cambridge University Press, 2018.
[41] T. Vu, R. Raich, Accelerating iterative hard thresholding for low-rank matrix completion via adaptive restart, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2917–2921.
[42] T. Zhang, Sparse recovery with orthogonal matching pursuit under RIP, IEEE Trans. Inf. Theory 57 (9) (2011) 6215–6221.