RESEARCH

Fast and low-memory compressive sensing algorithms for low tucker-rank tensor approximation from streamed measurements

Cullen Haselby[1] · Mark A. Iwen[1,2] · Deanna Needell[3] · Elizaveta Rebrova[4] · William Swartworth[3]

© The Author(s) 2025

Abstract

In this paper we consider the problem of recovering a low-rank Tucker approximation to a massive tensor based solely on structured random compressive measurements (i.e., a sketch). Crucially, the proposed random measurement ensembles are both designed to be compactly represented (i.e., low-memory), and can also be efficiently computed in one-pass over the tensor. Thus, the proposed compressive sensing approach may be used to produce a low-rank factorization of a huge tensor that is too large to store in memory with a total memory footprint on the order of the much smaller desired low-rank factorization. In addition, the compressive sensing recovery algorithm itself (which takes the compressive measurements as input, and then outputs a low-rank factorization) also runs in a time which principally depends only on the size of the sought factorization, making its runtime sub-linear in the size of the large tensor one is approximating. Finally, unlike prior works related to (streaming) algorithms for low-rank tensor approximation from such compressive measurements, we present a unified analysis of both Kronecker and Khatri-Rao structured measurement ensembles culminating in error guarantees comparing the error of our recovery algorithm's approximation of the input tensor to the best possible low-rank Tucker approximation error achievable for the tensor by any possible algorithm. We further include an empirical study of the proposed approach that verifies our theoretical findings and explores various trade-offs of parameters of interest.

Keywords Tensor · Tucker decomposition · Compressive sensing · Streaming

Mathematics Subject Classification (2010) 65F55

1 Introduction

With the rapid increase in data acquisition and data-driven applications comes the need for efficient methods to acquire, store, reconstruct, and analyze large-scale data.

Extended author information available on the last page of the article

Ⓐ Springer

In many settings, this data is multi-modal, making a tensor representation the most natural. These settings range from medical to communications to a widespread use in data science in general [1–5]. Although the tensor is a higher mode analogue of the matrix, linear algebraic and computational results for matrices generally do not extend beyond two modes. However, as with matrices, tensor data often has an implicit low-rank structure that can be utilized for efficient computation. In contrast to matrices, there are several non-equivalent notions of low-rankness for tensors. Such low-rank structures structures arise from factorizations such as Tucker, CANDECOMP/PARAFAC (CP), tubal, and tensor train [1].

Because of the large-scale nature of tensor data – both in the number of modes and the dimensionality of each mode – computationally efficient methods are critical for computing such factorizations, as well as recovering the data from compressed measurements. An ideal framework for this recovery should reduce overall memory requirements, involve measurement schemes with fast matrix-vector multiplies, apply to a wide range of low-rank decompositions, be robust to non-exact low-rankness, and provide a computationally tractable, provably accurate recovery scheme. The framework should apply to both static and streaming settings, the latter occurring when the tensor data is being updated sequentially over time (e.g. slice by slice, entry-wise, or rank-one updates) and when one wishes to save on space by only storing a compressed version of the data.

1.1 Prior work

A number of papers aim to obtain low-rank tensor decompositions in a fast an efficient way, often using randomization. Some of the Tucker-rank related works include [6–9]. Here, we assume that the tensor is received as part of a stream of data, i.e., the complete raw tensor is not stored and we minimize the number of visits to each tensor entry (to at most one or two times each, referring to the one-pass or two-pass algorithms, respectively). In the earlier work [10], the authors describe a variant of Tucker-Alternating Least Squares (aka Higher Order Orthogonal Iteration, multi-pass scenario) that employed TensorSketch to produce the necessary measurements. However, the quality of reconstruction for iterative schemes is sensitive to the initialization used for the factors and core, as well as requiring possibly many iterations to converge or overcome "swamps", a well documented issue with the ALS approach.

Herein we aim to recover a low-rank tensor from oblivious linear measurements such as those employed in compressive sensing tasks. Some of these measurements satisfy the Tensor Restricted Isometry Property (TRIP) and are amenable to provable recovery by iterative algorithms like the measurements and recovery procedures discussed in [11–13]. One of the main goals of this work is to make related measurement maps and recovery algorithms more memory-efficient by using more structured linear measurements. It is known to be hard to design such measurements that also satisfy the TRIP [13, 14]. As a result, in this work we employ direct recovery procedures that allow for error analysis that does not rely on the TRIP.

In the prior work of Hendrikx and De Lathauwer [15], the authors describe an algorithm that recovers a tensor from linear combinations of its entries where the

measurement operations themselves are constrained to be Kronecker-structured. In that work they give necessary and sufficient conditions for perfect recovery of exactly low-rank tensors. They also describe some relevant heuristics and provide empirical results for the performance of the recovery when the tensor or its measurements are subject to white noise. Additionally they demonstrate how to adapt their method to two other tensor formats: CP and tensor train.

In the work of Sun et al. [16], a comparable low Tucker-rank recovery procedure to [15] was described (Algorithm 4.1, Tucker Sketch). However, the measurement ensembles analyzed therein were both random and structurally less constrained, allowing for a tractable probabilistic analysis (i.e., the Gaussian measurements used to estimate factors in [16] were not constrained to be strictly modewise as in the Kronecker-structured case considered in [15]). As a result, quasi-optimal low-rank approximation guarantees for unstructured Gaussian linear measurements of the recovered tensor are proven in expectation in [16]. Furthermore, motivated by huge memory requirements of unstructured Gaussian measurement maps analyzed therein (which will generally take up as much storage as the data tensor itself), the authors also empirically investigate the performance of more structured measurement maps constructed via Khatri-Rao products of smaller Gaussian matrices. In those experiments they show that these Khatri-Rao structured measurements deliver nearly as good approximations in practice as the unstructured memory-hungry maps they analyzed theoretically, with the additional benefit of also requiring significantly less space to store. Prior to the work of [16, 17] applied similar ideas to develop a single-pass sketch for PCA in the two-mode (matrix) case, but also required a Gaussian sketch. While our focus is on tensors, we expect that our approach will also make the analysis of a wide variety of related one-pass PCA methods for 2-mode tensors, such as those discussed in [17], more tractable.

1.2 Contribution

In this work, we further study recovery from the Kronecker and Khatri-Rao measurement ensembles introduced in [15, 16]. Our analysis unifies and adds to both of these works in several ways. First, unlike in [15], our error analysis applies in the non-exact low-rank case, and quantifies how our low-rank approximation error bounds depend on the relevant parameters of the problem. Second, though the overarching structure of our theoretical analysis employs a similar strategy to that in [16], it is instead applied to more structured, memory efficient, and difficult to analyze Kronecker and Khatri-Rao structured measurement ensembles. The acquired measurements can then be used in a single unified recovery method (see Algorithm 1, which effectively matches those utilized in both [16] and [15]) to produce a low-rank factorization.

Among other differences from [16], we believe it is important to emphasize that the more structured measurement ensembles analyzed herein can operate on the tensor data in subtly different modewise manners than the measurements analyzed there. That is, our measurements yield several smaller tensors whose entries are obtained by matrix-vector operations involving measurement matrices operating on various fibers or slices of the tensor data. This can have significant practical advantages since neither the entire original tensor nor the entire measurement operator need to be kept in working mem-

ory in order to obtain such measurements. Additionally, besides applying the analysis to these more structured measurement ensembles, we remove the theoretical assumption in [16] that the entries of the sketching matrices are drawn independently from a Gaussian distribution, and instead rely on a more general Johnson-Lindenstrauss property-based analysis in order to derive with-high-probability recovery guarantees for more general measurement ensembles. The advantage of this is that there are several well known distributions of random matrices with, e.g., fast matrix-vector multiplies that are known to satisfy this property, and our alternate analysis makes it a straightforward exercise to derive measurement requirements and error bounds for measurement ensembles constructed from such different choices of sketching matrices.

Theorem 5 summarizes our main results. In order to state it, we will first describe some tensor concepts and operations related to the measurement process. The precise statements of the results are given in Theorem 39 (for generic measurement ensembles) and Theorems 41 and 43 (specialized for sub-gaussian measurement ensembles).

1.3 Tensor and measurement preliminaries

Here we recall terminology and set notation that will be useful in describing the tensor operations used in stating our main result and throughout this paper. For a more thorough introduction we refer the reader to, e.g., [1, 18].

Tensor order, fibers, and unfoldings The *order* of a tensor is its number of *modes*. That is, $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots n_d}$ is an order d tensor, or a d-mode tensor. *Mode-j fibers* are the tensor analogue of rows and columns in the matrix case. They are the vectors given by fixing all but one of the indices and varying the j-th coordinate. For example, the 3-mode tensor tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ will have $n_2 n_3$ mode-1 fibers $\mathbf{x}_{:,j,k} \in \mathbb{R}^{n_1}$ indexed by $j \in [n_2]$ and $k \in [n_3]$, $n_1 n_3$ mode-2 fibers $\mathbf{x}_{i,:,k} \in \mathbb{R}^{n_2}$ indexed by $i \in [n_1]$ and $k \in [n_3]$, and $n_1 n_2$ mode-3 fibers $\mathbf{x}_{i,j,:} \in \mathbb{R}^{n_3}$ indexed by $i \in [n_1]$ and $j \in [n_2]$. The *mode-j unfolding* of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots n_d}$ is an $n_j \times \prod_{k=1, k \neq j}^{d} n_k$ matrix, $X_{[j]}$, formed by arranging the mode-j fibers of $\mathcal{X}$ as the columns of $X_{[j]}$. The ordering of these columns is not important so long as it is consistent across calculations.

Inner product and norm The set of all d-mode tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ forms a vector space when equipped with component-wise addition and scalar multiplication. The *standard inner product* of two tensors $X, Y \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is

$$\langle \mathcal{X}, \mathcal{Y} \rangle := \sum_{j_1 \in [n_1], \ldots j_d \in [n_d]} \mathcal{X}_{j_1, \ldots, j_d} \mathcal{Y}_{j_1, \ldots, j_d}.$$

The standard tensor norm is then defined by this inner product as $\|\mathcal{X}\|_2 := \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$. Note that the standard tensor inner product and norm above correspond to the standard dot product and Euclidean norm as applied to vectorized tensors in $\mathbb{R}^N$, where $N = \prod_{j=1}^{d} n_j$.

Random variables It will also be useful to discuss probabilistic quantities. We denote the probability that a random variable Y belongs to a set E by $\mathbb{P}(Y \in E)$. We further

write the expectation of a random variable Y by $\mathbb{E}(Y)$, and refer to $\mathbb{E}(|Y|^p)$ as its p-th absolute moment. The p-th root of the p-th absolute moment of a random variable Y is also referred to as its L^p norm. It is denoted by

$$\|Y\|_{L^p} := (\mathbb{E}\,|Y|^p)^{1/p}$$

for all $p \in [1, \infty)$.

Definition 1 (Modewise product) The j-mode product of a d-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ with a matrix $\Phi \in \mathbb{R}^{m_j \times n_j}$ is denoted as $\mathcal{X} \times_j \Phi$, and defined componentwise as

$$\left(\mathcal{X} \times_j \Phi\right)_{i_1,\ldots,i_{j-1},\ell,i_{j+1},\ldots,i_d} = \sum_{i_j=1}^{n_j} \Phi_{\ell,i_j} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} \tag{1}$$

for $\ell \in [m_j]$ and $i_j \in [n_j]$. Equivalently, for the respective unfoldings, $\left(\mathcal{X} \times_j \Phi\right)_{[j]} = \Phi X_{[j]}$.

Additionally, we let $\circ$ denote the outer product of two tensors. That is, if $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $\mathcal{Y} \in \mathbb{R}^{m_1 \times \cdots \times m_f}$, then $\mathcal{X} \circ \mathcal{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_d \times m_1 \times \cdots \times m_f}$ with entries given by $(\mathcal{X} \circ \mathcal{Y})_{j_1,\ldots,j_d,i_1,\ldots,i_f} := \mathcal{X}_{j_1,\ldots,j_d} \mathcal{Y}_{i_1,\ldots,i_f}$.

1.3.1 The Tucker decomposition and low-rank approximation

Next, we define the Tucker decomposition of tensors which decomposes a tensor into a core tensor that is multiplied by a factor matrix along each of its modes. Crucially, this decomposition relates to the proposed modewise measurement operations in a convenient fashion as we detail in Section 3.

Definition 2 (Tucker decomposition) Let $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. We say that $\mathcal{X}$ has a Tucker decomposition of rank $\mathbf{r} = (r_1, \ldots, r_d)$ if there exists $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$, and $U_j \in \mathbb{R}^{n_j \times r_j}$ with orthonormal columns for all $j \in [d]$, so that

$$\mathcal{X} = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d = \sum_{\mathbf{i} \in \{(i_1,\ldots,i_d)\,|\,i_j \in [r_j]\,\forall j \in [d]\}} \mathcal{G}_{\mathbf{i}} \cdot \mathbf{u}_{i_1}^{(1)} \circ \mathbf{u}_{i_2}^{(2)} \circ \cdots \circ \mathbf{u}_{i_d}^{(d)},$$
$$\tag{2}$$

where $\mathbf{u}_{i_j}^{(j)}$ denotes the i_j-th column of U_j, and $\mathcal{G}_{\mathbf{i}} := \mathcal{G}_{i_1,\ldots,i_d} \in \mathbb{R}$. We will also use the shorthand $\mathcal{X} = [\![\mathcal{G}, U_1, \ldots, U_d]\!] := \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d$ when referring to the rank-$\mathbf{r}$ Tucker decomposition of a given rank-$\mathbf{r}$ tensor $\mathcal{X}$.

There is an equivalent formulation to (2) using unfoldings of $\mathcal{X}$ and of $\mathcal{G}$ in terms of matrix-matrix products:

$$X_{[j]} = U_j G_{[j]} \left(U_d \otimes \cdots \otimes U_{j+1} \otimes U_{j-1} \otimes \cdots \otimes U_1\right)^T, \tag{3}$$

where $\otimes$ denotes the standard Kronecker matrix product (see (6) and, e.g., [19]).

Given an arbitrary tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ of unknown Tucker rank it is often of interest to compute a good rank-$\mathbf{r}$ approximation of $\mathcal{X}$. In fact, an optimal Tucker rank-$\mathbf{r}$ minimizer $[\![\mathcal{X}]\!]_{\mathbf{r}}^{\text{opt}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ satisfying

$$\left\| \mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}^{\text{opt}} \right\|_2 = \inf_{\substack{\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d}, \\ U_j \in \mathbb{R}^{n_j \times r_j}}} \| \mathcal{X} - [\![\mathcal{G}, U_1, \ldots, U_d]\!] \|_2 \qquad (4)$$

always exists (see, e.g., Theorems 10.3 and Theorem 10.8 in [20]). However, computing such an $[\![\mathcal{X}]\!]_{\mathbf{r}}^{\text{opt}}$ (which is not unique) is generally a challenging task that is accomplished only approximately via iterative techniques (see, e.g., [1]). As a result, in such situations one usually seeks to instead compute a *quasi-optimal* rank-$\mathbf{r}$ tensor $\tilde{\mathcal{X}}$ satisfying

$$\left\| \mathcal{X} - \tilde{\mathcal{X}} \right\|_2 \leqslant C \inf_{\substack{\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d}, \\ U_j \in \mathbb{R}^{n_j \times r_j}}} \| \mathcal{X} - [\![\mathcal{G}, U_1, \ldots, U_d]\!] \|_2 , \qquad (5)$$

where $C \in \mathbb{R}^+$ is a positive constant independent of $\mathcal{X}$.

The next lemma can be used to demonstrate that one can recover a quasi-optimal rank-$\mathbf{r}$ approximation of an arbitrary tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ by simply computing the left singular vectors of $X_{[j]}$ for all $j \in [d]$. That is, the resulting modewise-defined rank-$\mathbf{r}$ projection (playing the role of $\tilde{\mathcal{X}}$ in (5) above) can be shown to be quasi-optimal by first bounding the error between $X_{[j]}$ and its best rank-r_j projection from above by the optimal Tucker approximation error (4) $\forall j \in [d]$. The error between $X_{[j]}$ and its best rank-r_j projection is, in turn, equal to the sum of the squares of the trailing singular values of $X_{[j]}$ $\forall j \in [d]$ by the Eckart-Young Theorem. The following related lemma is a variant of [20, Theorem 10.3] which we prove here for the sake of completeness.

Lemma 3 *Let $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, denote the ith-largest singular value of $X_{[j]}$ by $\sigma_i \left(X_{[j]} \right)$, and define $\Delta_{r,j} := \sum_{i=r+1}^{\tilde{n}_j} \sigma_i \left(X_{[j]} \right)^2$ for all $j \in [d]$ and $r \in \left[\tilde{n}_j := \min \left\{ n_j, \prod_{j \neq k} n_k \right\} \right]$. Fix $\mathbf{r} \in [\tilde{n}_1] \times \cdots \times [\tilde{n}_d]$ and suppose that $[\![\mathcal{X}]\!]_{\mathbf{r}}^{opt}$ satisfies (4). Then,*

$$\sum_{j=1}^d \Delta_{r_j, j} \leqslant d \left\| \mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}^{opt} \right\|_2^2 .$$

Proof Let $P_j \in \mathbb{R}^{n_j \times n_j}$ be a rank r_j orthogonal projection matrix satisfying $\left\| X_{[j]} - P_j X_{[j]} \right\|_F^2 = \Delta_{r_j, j} \leqslant \left\| X_{[j]} - Y \right\|_F^2$ for all rank at most r_j matrices Y.[1] Noting that $Y = \left([\![\mathcal{X}]\!]_{\mathbf{r}}^{opt} \right)_{[j]}$ will be a rank at most r_j matrix by (3), we then have that

$$\Delta_{r_j, j} \leqslant \left\| X_{[j]} - \left([\![\mathcal{X}]\!]_{\mathbf{r}}^{opt} \right)_{[j]} \right\|_F^2 = \left\| \mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}^{opt} \right\|_2^2 \quad \forall j \in [d].$$

---

[1] The Eckart-Young Theorem guarantees that we can compute such a P_j from the left singular values of $X_{[j]}$.

Summing over the d values of j above now finishes the proof. $\qquad\square$

Note, in order to simplify notation moving forward, we will assume that $n_j = n$ and $r_j = r$ for all $j \in [d]$; that is our tensors will have equal side lengths and the rank will likewise be the same in every mode, thus suppressing the need for additional level of sub-scripting.

1.4 Kronecker products, Khatri-Rao products, and leave-one-out modewise measurement notation and examples

We now will describe and provide some notation for the measurement operators used throughout the rest of this paper. The general concept we are able to specialize to several cases is the following:

Definition 4 (Leave-One-Out Measurements) Given a d-mode tensor $\mathcal{X}$ with side-lengths n, leave-one-out measurements are the result of reducing all but one mode using linear operations on the unfolding of the tensor. That is,

$$B_j = \Omega_{(j,j)} X_{[j]} \Omega_{-j}^T$$

are leave one out measurements whenever $\Omega_{(j,j)} \in \mathbb{R}^{n \times n}$ and full-rank, $\Omega_{-j} \in \mathbb{R}^{m' \times n^{d-1}}$ where $m' \leqslant n^{d-1}$.

Any measurement process that can be equivalently described as left multiplication of an unfolded tensor by a full-rank square matrix and right multiplied by some other linear operator that reduces all other modes is a leave-one measurement process. We will consider three specific cases for how to construct the overall measurement operator Ω_{-j} in Definition 4: Kronecker structured, Khatri-Rao structured and unstructured measurement operators.

We will need to recall some matrix products. The *Kronecker product* of matrices $A \in \mathbb{R}^{m_1 \times n_1}$ and $B \in \mathbb{R}^{m_2 \times n_2}$ is the matrix $A \otimes B \in \mathbb{R}^{m_1 m_2 \times n_1 n_2}$ and is defined by

$$A \otimes B := \begin{bmatrix} a_{1,1}B & a_{1,2}B & \dots & a_{1,n_1}B \\ a_{2,1}B & a_{2,2}B & \dots & a_{2,n_1}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m_1,1}B & a_{m_1,2}B & \dots & a_{m_1,n_1}B \end{bmatrix}. \tag{6}$$

The *Khatri-Rao product* of two matrices is the matrix that results from computing the Kronecker product of their matching columns. That is, for $A \in \mathbb{R}^{m_1 \times n}$ and $B \in \mathbb{R}^{m_2 \times n}$, their Khatri-Rao product is the matrix $A \odot B \in \mathbb{R}^{m_1 m_2 \times n}$ defined by

$$A \odot B := \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \mathbf{a}_2 \otimes \mathbf{b}_2 & \dots & \mathbf{a}_n \otimes \mathbf{b}_n \end{bmatrix}.$$

In this paper, we also use a so-called row-wise Khatri-Rao product (sometimes called the *face-splitting product*) of $A \in \mathbb{R}^{m \times n_1}$ and $B \in \mathbb{R}^{m \times n_2}$, denoted by $A \bullet B \in$

$\mathbb{R}^{m \times n_1 n_2}$, and defined by $(A \bullet B)^T := A^T \odot B^T$. Given the close relationship between the Khatri-Rao and face-splitting products, we will refer to both of them as being "Khatri-Rao structured" below.

1.4.1 Kronecker structured measurements

Kronecker structured leave-one-out measurements are constructed by taking the Kronecker product of several component matrices, in addition to one square measurement matrix of full-rank to be applied to the mode which is uncompressed (which may simply be the identity). We will require leave one out measurements for each mode. Collectively the matrices needed to define this ensemble will form a set $\{\Omega_{(i,j)}\}_{i,j \in [d]}$ where $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ when $j \neq i$, and where $\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ for all $i \in [d]$.

Our measurements then are

$$\mathcal{B}_j := \mathcal{X} \bigtimes_{i=1}^{d} \Omega_{(j,i)} \tag{7}$$

for all $j \in [d]$ using the matrices $\{\Omega_{(j,i)}\}_{i=1}^{d}$. Crucially, we can identify this tensor of measurements $\mathcal{B}_j$ with a flattened version which makes it clear that Kronecker structured modewise measurements do define leave-one-out measurements conforming to Definition 4.

$$B_j = \Omega_{(j,j)} X_{[j]} \Omega_{-j}^T = \Omega_{(j,j)} X_{[j]} (\Omega_{(j,1)} \otimes \Omega_{(j,2)} \otimes \ldots \Omega_{j,j-1} \otimes \Omega_{(j,j+1)} \otimes \cdots \otimes \Omega_{(j,d)})^T \tag{8}$$

Conceptually, the Kronecker products of matrices can be rewritten in terms of matrix products of auxiliary matrices. This will be useful if we wish to examine via, e.g., (3), how several modewise products change the properties of the resulting factor matrices U_j in the Tucker decomposition of a given tensor. As an illustration, consider the case of a three mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. Let $\mathbf{x} \in \mathbb{R}^{n_1 n_2 n_3}$ denote the vectorization of $\mathcal{X}$, and further suppose that $\Omega_j \in \mathbb{R}^{m_j \times n_j}$ for $j = 1, 2, 3$ are three measurement matrices used to produce modewise measurements of $\mathcal{X}$ given by $\mathcal{X} \times_1 \Omega_1 \times_2 \Omega_2 \times_3 \Omega_3$. Allowing for an additional reshaping, we can identify these three modewise operations equivalently with a single matrix-vector product using a variant of (3). That is, one can see that vec$(\mathcal{X} \times_1 \Omega_1 \times_2 \Omega_2 \times_3 \Omega_3) = \Omega \mathbf{x}$ where $\Omega = \Omega_3 \otimes \Omega_2 \otimes \Omega_1 \in \mathbb{R}^{m_1 m_2 m_3 \times n_1 n_2 n_3}$. Let I_n denote the $n \times n$ identity matrix. The mixed-product property of the Kronecker product can now be used to further show that in fact $\Omega = \tilde{\Omega}_3 \tilde{\Omega}_2 \tilde{\Omega}_1$, where

$$\tilde{\Omega}_1 = I_{n_3} \otimes I_{n_2} \otimes \Omega_1 \in \mathbb{R}^{m_1 n_2 n_3 \times n_1 n_2 n_3},$$
$$\tilde{\Omega}_2 = I_{n_3} \otimes \Omega_2 \otimes I_{m_1} \in \mathbb{R}^{m_1 m_2 n_3 \times m_1 n_2 n_3},$$
$$\tilde{\Omega}_3 = \Omega_3 \otimes I_{m_2} \otimes I_{m_1} \in \mathbb{R}^{m_1 m_2 m_3 \times m_1 m_2 n_3}.$$

Hence, vec$(\mathcal{X} \times_1 \Omega_1 \times_2 \Omega_2 \times_3 \Omega_3) = \tilde{\Omega}_3 \tilde{\Omega}_2 \tilde{\Omega}_1vec(\mathcal{X})$. This example can be easily be extended to any number of modes.

Computing Kronecker Structured Measurements in One Pass

Given $(i_1, \ldots, i_d) \in [n]^d$ let $\mathcal{E}^{(i_1,\ldots,i_d)} \in \mathbb{R}^{n \times \cdots \times n}$ be the canonical basis tensor defined entrywise by

$$\mathcal{E}^{(i_1,\ldots,i_d)}_{j_1,\ldots,j_d} = \begin{cases} 1 & \text{if } (j_1, \ldots, j_d) = (i_1, \ldots, i_d) \\ 0 & \text{else} \end{cases}.$$

Fix $j \in [d]$ and let be $z_{-j} = 1 + \sum_{k=1}^{j-1}(i_k - 1)n^{d-k-1} + \sum_{k=j+1}^{d}(i_k - 1)n^{d-k}$. The mode-$j$ unfolding of $\mathcal{E}^{(i_1,\ldots,i_j,\ldots,i_d)}$ can then be given by $E^{(i_1,\ldots,i_d)}_{[j]} = E^{(i_j,z_{-j})} \in \mathbb{R}^{n \times n^{d-1}}$ which can be written entrywise by

$$E^{(i_1,\ldots,i_d)}_{[j]} = E^{(i_j,z_{-j})} = \begin{cases} 1 & \text{if } (h, \ell) = (i_j, z_{-j}) \\ 0 & \text{else} \end{cases}.$$

For any given tensor $\mathcal{X} \in \mathbb{R}^{n \times \cdots \times n}$ the flattened Kronecker measurement (7) can now be computed via

$$
\begin{aligned}
B_j &= \Omega_{(j,j)} X_{[j]} \Omega_{-j}^T = \Omega_{(j,j)} \left(\sum_{(i_1,\ldots,i_d) \in [n]^d} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} E^{(i_j,z_{-j})} \right) \Omega_{-j}^T \\
&= \sum_{(i_1,\ldots,i_d) \in [n]^d} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} \Omega_{(j,j)} E^{(i_j,z_{-j})} \Omega_{-j}^T \qquad (9) \\
&= \sum_{(i_1,\ldots,i_d) \in [n]^d} \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} \left(\Omega_{(j,j)} \right)_{:,i_j} \left(\left(\Omega_{-j} \right)_{:,z_{-j}} \right)^T.
\end{aligned}
$$

Note that (9) allows each Kronecker measurement B_j to be computed in just one pass over the entries of $\mathcal{X}$ by sequentially summing weighted outer products of columns of $\Omega_{(j,j)}$ with columns of Ω_{-j}. In a streaming setting (e.g., under the "turnstile model" [21]) the influence of entrywise increments to $\mathcal{X}$ on each B_j measurement can be tracked analogously. See Algorithm 2 for an outline of this sketching procedure.

Another common choice one might make in passing through the entries of $\mathcal{X}$ once would be to load it as a series of sub-tensors. For example, suppose for simplicity that $d = 3$ and that we load $\mathcal{X} \in \mathbb{R}^{n \times n \times n}$ as a series of n matrices $\{\mathcal{X}_{:,:,k}\}_{k \in [n]} \subset \mathbb{R}^{n \times n}$. Fix $j \in [3]$. In this setting the Kronecker structured measurements (7) can be partially computed easily on each submatrix since the partial Kronecker measurements

$$\mathcal{B}'_j := \mathcal{X} \times_1 \Omega_{(j,1)} \times_2 \Omega_{(j,2)}$$

will satisfy $(\mathcal{B}'_j)_{:,:,k} = \Omega_{(j,1)}(\mathcal{X}_{:,:,k})\Omega_{(j,2)}^T$ for all $k \in [n]$. Hence, if we define the new tensor $\tilde{\mathcal{E}}^{(i_1,i_2;i_3)}$ entrywise via

$$\tilde{\mathcal{E}}^{(i_1,i_2;i_3)}_{j_1,j_2,j_3} = \begin{cases} \left(\Omega_{(j,3)}\right)_{j_3,i_3} & \text{if } (j_1, j_2) = (i_1, i_2) \\ 0 & \text{else} \end{cases},$$

then we can see that

$$\mathcal{B}_j = \sum_k \left[\sum_{i_1,i_2} \left(\Omega_{(j,1)}(\mathcal{X}_{:,:,k})\Omega_{(j,2)}^T \right)_{i_1,i_2} \tilde{\mathcal{E}}^{(i_1,i_2;k)} \right]. \qquad (10)$$

Note that (10) allows each Kronecker structured measurement $\mathcal{B}_j$ to be incrementally computed in just one pass over these n submatrices of $\mathcal{X}$. Of course, many other one-pass strategies for computing these measurements also exist depending on how one wants to/can read through $\mathcal{X}$.

1.4.2 Khatri-Rao structured measurements

Leave-one-out measurement ensembles which are Khatri-Rao products of smaller maps are also possible and have been considered before in works such [16]. That is,

$$\begin{aligned} B_j &= \Omega_{(j,j)} X_{[j]} \Omega_{-j}^T \\ &= \Omega_{(j,j)} X_{[j]} \left(\Omega_{(j,1)}^T \odot \Omega_{(j,2)}^T \odot \ldots \Omega_{(j,j-1)}^T \odot \Omega_{(j,j+1)}^T \cdots \odot \Omega_{(j,d)}^T \right)^T \\ &= \Omega_{(j,j)} X_{[j]} \left(\Omega_{(j,1)} \bullet \Omega_{(j,2)} \bullet \ldots \Omega_{(j,j-1)}^T \bullet \Omega_{(j,j+1)} \ldots \bullet \Omega_{(j,d)} \right)^T \qquad (11) \end{aligned}$$

where $\Omega_{(j,j)}$ is a full-rank square matrix, and $\Omega_{(j,i)} \in \mathbb{R}^{m \times n}$ for $j \neq i$. Note that in this case we can consider $\Omega_{-j} \in \mathbb{R}^{m \times n^{d-1}}$ as sketching sub-tensors of size n^{d-1} to size m (as opposed to m^{d-1} as was done previously with Kronecker structured measurement maps). See Algorithm 3 for a summary of the required sketches.

Computing Khatri-Rao Structured Measurements in One Pass
Given they are linear measurements, Khatri-Rao structured measurements can also be updated in flattened form according to (9). For additional implementational details on computing such measurements in one pass we refer the interested reader to the relevant sections on the "Tensor Random Projection (TRP)" in [16]. Herein we primarily focus instead on the theoretical analysis of such measurements (see, e.g., §3.3).

1.4.3 Unstructured measurements

In [22], a type of leave-one-out measurements are described and analyzed, however in that work their measurement ensembles are not structured, but instead are Gaussian matrices where $\Omega_{-j} \in \mathbb{R}^{m \times n^{d-1}}$ has entries all drawn independently and is right multiplied by an unfolding of the tensor:

$$B_j = \Omega_{(j,j)} X_{[j]} \Omega_{-j}^T$$

where again $\Omega_{(j,j)}$ is some square, full-rank matrix.

As an important observation, the storage requirements for the sketching matrix itself in this case is large, comparable in size to the data itself, which can be undesirable. Furthermore, the error analysis in [22] conducted on this type of measurement when the unstructured matrix Ω_{-j} has i.i.d. Gaussian entries relies on probabilistic bounds known for Gaussian matrices. Comparable bounds for other distributions, or for matrices constructed using Kronecker or Khatri-Rhao products are not covered in the analysis. Note, Kronecker or Khatri-Rao structured measurements alleviate to a large degree the storage problem associated with $\Omega_{-j} \in \mathbb{R}^{m \times n^{d-1}}$ when compared to in the unstructured case since the entire sketching matrix does not need to be maintained in memory, but rather just constituent components $\Omega_{(j,i)} \in \mathbb{R}^{m \times n}$ can be stored, and the action of the Kronecker or Khatri-Rao product on the tensor can computed as part of the sketching phase. This does incur additional operations in the sketching phase, and in our runtime analysis and numerical experiments we detail the trade-off between space and run-time for these various choices of measurement operators.

Computing Unstructured Measurements in One Pass

Again, unstructured measurements can also be updated in flattened form using, e.g., (9). Of course, the most efficient means of doing so will vary depending on the situation/application.

1.4.4 Core measurements

Regardless of which type leave-one-measurements are used for each of the modes, in order to recover the full factorization using only a single access to the data $\mathcal{X}$, we will require an additional set of measurements for use in estimating the core. These are in all cases, modewise, and can compress all modes.

$$\mathcal{B}_c := \mathcal{X} \bigtimes_{j=1}^{d} \Phi_j \tag{12}$$

using the matrices $\{\Phi_j\}_{j=1}^{d}$. See Algorithm 4 for a summary of the required core sketches.

All together then, these $d + 1$ measurement tensors, d of the leave-one-out type and one of the type in (12) will be used to recover the parts of the original full tensor. Figure 1 is a schematic rendition of the overall measurement procedure in the three mode case ($d = 3$).

Note that the n^d entries of the original tensor $\mathcal{X}$ are compressed into $dnm^{d-1} + m_c^d$ entries of the $d+1$ total different measurement tensors; one leave-one-out measurement for each of the d modes as well as the one measurement tensor for use in estimating the core. Recovery naturally will require the storage of the measurement operators in some fashion along with the measurement tensors. As dense matrices, the measurement operators collectively have $d((d-1)mn + n^2) + dm_c n$ total entries. However the number of random variables required to generate these matrices may be significantly
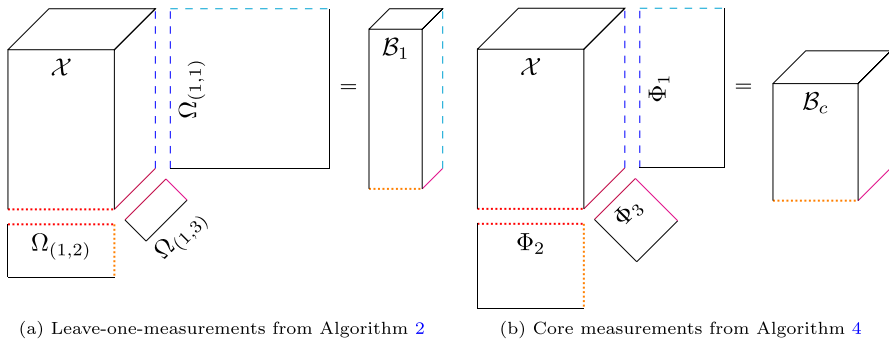
(a) Leave-one-measurements from Algorithm 2   (b) Core measurements from Algorithm 4

Fig. 1 Schematic for the two types of measurement tensors used in recovery Algorithm 9 for a three mode tensor. Measurement tensors $\mathcal{B}_i$ of the type shown in (a) for $i = 1$ will be used to estimate the factors of the Tucker decomposition, and the measurement tensor $\mathcal{B}_c$ in (b) will be used to estimate the core of the Tucker decomposition

fewer depending on the method employed. For example, when using sub-sampled Fourier transforms, the number of random bits required to construct an ensemble is linear in n. Additionally, some choices of measurement matrices allow for near-linear time matrix-vector multiplication. This is often accomplished by choosing matrices that exploit either the Fast-Fourier transform, or fast Hadamard multiplication, and will allow us to compress our tensors faster than with Gaussian measurements for example.

As we detail, applying the measurement operators is asymptotically the most computationally intensive part of the algorithm, and thus in settings where it is useful to economize the computational effort to obtain measurements these matrices have significant advantages. Whichever choice for type of measurements are used, however, the efficiency in terms of the run-time of the recovery algorithm is largely dependent on the ratios m/n and m_c/n.

Computing Core Measurements in One Pass
Core measurements are another form of Kronecker structured measurements. Given this, they can be computed in one pass using all of the same techniques already discussed in Section 1.4.1.

1.4.5 The canonical one-pass recovery algorithm

We can now state Algorithm 1 as the canonical algorithm for one-pass recovery the tensor using leave-one-out measurements. The algorithm outputs an estimate in factored form $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$, and we say one-pass to emphasize that after measuring the tensor $\mathcal{X}$, no additional access to the data is required to obtain the estimate $\mathcal{X}_1$. The inputs to the algorithm are leave-one-out measurements B_i for each of the modes of any type, as well as measurements $\mathcal{B}_c$ for use in recovering the core, some of the related measurement operators $\Omega_{(i,i)}$, Φ_i, and a target rank **r** parameter. Algorithm 1 consists of two loops. The first loop recovers the factor matrices Q_i in the Tucker

factorization. The second loop depends on the output of the first loop and computes the core $\mathcal{H}$ of the Tucker factorization.

Note that within the pseudo-code for Algorithm 1 the "unfold" function in the body of Algorithm 1 takes a tensor and flattens it into the given shape by arranging the specified mode's fibers as the columns, e.g. $H \leftarrow \text{unfold}(\mathcal{B}_c, m_c \times m_c^{d-1}, \text{mode} = 1)$ is equivalent to $H \leftarrow (\mathcal{B}_c)_{[1]}$ using the notation of Section 1.3, the matrix that is obtained by flattening the d-mode tensor $\mathcal{B}_c$ along mode-1. The "fold" function is the inverse of "unfold", and takes a matrix that is an unfolding along the specified mode of a tensor and reshapes it into a tensor with the specified and compatible dimensions, e.g. $\mathcal{B} \leftarrow \text{fold}(H, r \underbrace{\times m_c \times \cdots \times m_c}_{d-1}, \text{mode} = 1)$ takes the matrix H with dimensions $r \times m_c^{d-1}$ and reshapes it into a d mode tensor where the first mode is of length r and the remaining $d - 1$ modes are of length m_c.

In the scenario where it is possible to access the original tensor $\mathcal{X}$ twice (we refer to this as the two-pass scenario), we can compute the optimal core $\mathcal{G}$ given our estimated factor matrices Q_i to obtain an estimate $\mathcal{X}_2 = [\![\mathcal{G}, Q_1, \ldots, Q_d]\!]$. See Algorithm 10 and Algorithm 12 in Appendix B for the detailed formulation of the two-pass recovery procedure.

With the measurement operators described and the canonical algorithm outlined, we can now state a representative main result. The full unabridged results are found in Section 3.5.

Theorem 5 (A Summary Main Result) *Let $\mathcal{X}$ be a d-mode tensor of side length n, $\epsilon \in (0, 1)$, $\delta \in (0, \frac{1}{3})$, and $r \in [n] := \{1, \ldots, n\}$. Denote an optimal rank $\mathbf{r} = (r, \ldots, r) \in \mathbb{R}^d$ ($d \geqslant 2$) Tucker approximation of $\mathcal{X}$ by $[\![\mathcal{X}]\!]_{\mathbf{r}}^{\text{opt}}$. There exists a one-pass recovery algorithm (see Algorithm 1) that outputs a Tucker factorization $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ of $\mathcal{X}$ from leave-one-out linear measurements that will satisfy*

$$\|\mathcal{X} - \mathcal{X}_1\|_2 \leqslant (1 + e^\epsilon)\sqrt{\frac{d(1 + \epsilon)}{1 - \epsilon}} \left\|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}^{\text{opt}}\right\|_2 \tag{13}$$

with probability at least $1 - \delta$. The total number of linear measurements the algorithm will use is bounded above by $m'(r, d, \epsilon, n, \delta) = \left[nd + \frac{Crd^2}{\epsilon^2} \ln\left(\frac{dn^d}{\delta}\right)\right] \left(\frac{Crd^2}{\epsilon^2} \ln\left(\frac{dn^d}{\delta}\right)\right)^{d-1}$, where $C > 0$ is an absolute constant. Furthermore, the recovery algorithm (Algorithm 1) runs in time $o(n^d)$ for large $n \gg r$.

Remark 6 More generally and more exactly, the number of measurements is always $ndm^{d-1} + m_c^d$ where m and m_c are as in Theorem 41 for Kronecker structured measurements, or $d\tilde{m}n + m_c^d$ where $\tilde{m}$ and m_c are as in Theorem 43 in the case of Khatri-Rao structured measurements.

Proof Combine Theorem 41 (or Theorem 43) with Theorem 42 and Lemma 3. □

Bounds that depend on the total size of the tensor are typical in similar compressive scenarios. Our bound by contrast, although it is not tight, depends on the number of

---

Algorithm 1 One Pass HOSVD Recovery from Leave-One-Out Measurements.

input :
$B_i \in \mathbb{R}^{n \times m^{d-1}}$ for $i \in [d]$ leave-one-out measurements
$\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ for $i \in [d]$ full-rank sensing matrix for uncompressed mode
$\mathcal{B}_c$ a d mode tensor of measurements with side lengths m_c
$\Phi_i \in \mathbb{R}^{m_c \times n}$ for $i \in [d]$ measurement matrices for core measurements
$\mathbf{r} = (r, r, \ldots, r)$ the rank of the HOSVD
output: $\hat{\mathcal{X}} = [\![\mathcal{H}, Q_1, \ldots Q_d]\!]$
\# Factor matrix recovery
for $i \in [d]$ **do**
  \# Solve $n \times n$ linear system
  Solve $\Omega_{(i,i)} F_i = B_i$ for F_i
  \# Compute SVD and keep the r leading singular
    vectors
  $U, \Sigma, V^T \leftarrow \text{SVD}(F_i)$
  $Q_i \leftarrow U_{:,:r}$
end
\# Core recovery
for $i \in [d]$ **do**
  \# unfold measurements, mode-i fibers are columns,
    size $m_c \times r^{(i-1)} m_c^{d-1-(i-1)}$
  $H \leftarrow \text{unfold}(\mathcal{B}_c, m_c \times r^{(i-1)} m_c^{d-1-(i-1)}, \text{mode} = i)$
  \# Undo the mode-i measurement operator and factor's
    action by finding least square solution to $m_c \times r$
    over-determined linear system
  Solve $\Phi_i Q_i H_{\text{new}} = H$ for H_{new}
  \# reshape the flattened partially solved core into
    a tensor
  $\mathcal{B}_c \leftarrow \text{fold}(H_{\text{new}}, \underbrace{r \times r \cdots \times r}_{i} \underbrace{\times m_c \times \cdots \times m_c}_{d-i}, \text{mode} = i)$
  \# Each iteration $m_c \to r$ in ith mode
end
$\mathcal{H} \leftarrow \mathcal{B}_c$

---

degrees of freedom of the sketched tensor and not its full size. Furthermore, in our numerical experiments we demonstrate some of the main trade-offs involved when choosing between Kronecker or Khatri-Rao constructed sketches in terms of approximation error and performance.

Remark 7 Note that Theorem 5 guarantees that Algorithm 1 will recover an exact Tucker factorization of any rank $\mathbf{r}$ tensor with high probability since the right-hand side of (13) will be 0 in such cases. Given this, one can see that Algorithm 1 must use $\Omega(n)$-measurements. That said, if n itself is prohibitively large one could certainly imagine, e.g., replacing the square $\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ matrices in our Kronecker structured

measurements with something else like, e.g., an $m \times n$ compressive sensing matrix instead. Doing so for all $i \in [d]$ would then indeed make m' scale poly-logarithmically in n instead of linearly in n. However, doing so would also mean that one could at best, e.g., produce a sparse approximation to an exact Tucker factorization (instead of an exact factorization itself).

2 Preliminaries

The following preliminaries will provide the necessary definitions and basic results needed to conduct our probabilistic error analysis of the various leave-one-out measurements and Algorithm 1. We will rely heavily on some definitions in randomized numerical linear algebra and compressive sensing.

2.1 Randomized numerical linear algebra preliminaries

Definition 8 ((ϵ, δ, p)**-Johnson-Lindenstrauss (JL) property**) Let $\epsilon > 0, \delta \in (0, 1)$, and $p \in \mathbb{N}$. A random matrix $\Omega \in \mathbb{R}^{m \times n}$ has the (ϵ, δ, p)-Johnson-Lindenstrauss (JL) property[2] for an arbitrary set $S \subset \mathbb{R}^n$ with cardinality at most p if it satisfies

$$(1 - \epsilon) \|\mathbf{x}\|_2^2 \leqslant \|\Omega \mathbf{x}\|_2^2 \leqslant (1 + \epsilon) \|\mathbf{x}\|_2^2 \quad \text{for all } \mathbf{x} \in S \tag{14}$$

with probability at least $1 - \delta$.

For all random matrix distributions discussed in this work, only an upper-bound on the cardinality of the set S appearing in Definition 14 is required in order to ensure with high probability that a given realization satisfies (14). No other property of the set S to be embedded will be required to define, generate, or apply Ω. Such random matrices are referred to as *data oblivious*, or simply *oblivious*. The following variant of the famous Johnson-Lindenstrauss Lemma [23] demonstrates the existence of random matrices with the oblivious JL property.

Theorem 9 (sub-gaussian random matrices have the JL property) *Let $S \subset \mathbb{R}^n$ be an arbitrary finite subset of $\mathbb{R}^n$. Let $\delta, \epsilon \in (0, 1)$. Finally, let $\Omega \in \mathbb{R}^{m \times n}$ be a matrix with independent, mean zero, variance m^{-1}, sub-gaussian entries all admitting the same parameter $c \in \mathbb{R}^+$.[3] Then*

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leqslant \|\Omega \mathbf{x}\|_2^2 \leqslant (1 + \epsilon)\|\mathbf{x}\|_2^2$$

will hold simultaneously for all $\mathbf{x} \in S$ with probability at least $1 - \delta$, provided that

$$m \geqslant \frac{C}{\epsilon^2} \ln \left(\frac{|S|}{\delta} \right), \tag{15}$$

---

[2] Formally, a distribution over $m \times n$ matrices has the JL property if a matrix selected from this distribution satisfies (14) with probability at least $1 - \delta$. For brevity, here and in the next similar cases, the term "random matrix" will refer to a distribution over matrices.

[3] See, e.g., Remark 7.25 in [24] for details regarding sub-gaussian random variables are their parameters.

where $C \leqslant 8c(16c+1)$ is an absolute constant that only depends on the sub-gaussian parameter c.

Proof See, e.g., Lemma 9.35 in [24]. □

Next, we define a similar property for an infinite, yet rank constrained, set.

2.1.1 Oblivious Subspace Embedding (OSE) results

In this section we will review so-called Oblivious Subspace Embedding property, and describe how to construct them from any random matrix distribution with the JL property.

Definition 10 [(ϵ, δ, r)-**OSE property**] Let $\epsilon > 0$, $\delta \in (0, 1)$, and $r \in \mathbb{N}$. Fix an arbitrary rank r matrix $A \in \mathbb{R}^{n \times N}$. A random matrix $\Omega \in \mathbb{R}^{m \times n}$ has the (ϵ, δ, r)-Oblivious Subspace Embedding (OSE) property for (the column space of) A if it satisfies

$$(1 - \epsilon) \|A\mathbf{x}\|_2^2 \leqslant \|\Omega A\mathbf{x}\|_2^2 \leqslant (1 + \epsilon) \|A\mathbf{x}\|_2^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^N \qquad (16)$$

with probability at least $1 - \delta$.

Note that if $Q \in \mathbb{R}^{n \times r}$ is an orthonormal basis for the column space of a rank r matrix $A \in \mathbb{R}^{n \times N}$, then the images of A and Q coincide, i.e. $\{Q\mathbf{y} \mid \forall \mathbf{y} \in \mathbb{R}^r\} = \{A\mathbf{x} \mid \forall \mathbf{x} \in \mathbb{R}^N\}$. Furthermore, since Q has orthonormal columns so that $\|Q\mathbf{y}\|_2 = \|\mathbf{y}\|_2$ holds, (16) is equivalent to

$$(1 - \epsilon) \|\mathbf{y}\|_2^2 \leqslant \|\Omega Q\mathbf{y}\|_2^2 \leqslant (1 + \epsilon) \|\mathbf{y}\|_2^2 \quad \text{for all } \mathbf{y} \in \mathbb{R}^r \qquad (17)$$

holding in this case. Below we will use the equivalence of (16) and (17) often by regularly constructing matrices satisfying (16) by instead constructing matrices satisfying (17).

The next result shows that random matrices with the JL property also have the OSE property. It is a standard result in the compressive sensing and randomized numerical linear algebra literature (see, e.g., [25, Lemma 5.1] or [26, Lemma 10]) that can be proven using an ϵ-cover of the appropriate column space.

Lemma 11 (Subspace embeddings from finite embeddings via a cover) *Fix $\epsilon \in (0, 1)$. Let $\mathcal{L}_\mathcal{B}^r \subset \mathbb{R}^n$ be the r-dimensional subspace of $\mathbb{R}^n$ spanned by an orthonormal basis $\mathcal{B}$, and define $S_\mathcal{B}^r := \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \mid \mathbf{x} \in \mathcal{L}_\mathcal{B}^r \setminus \{\mathbf{0}\} \right\}$. Furthermore, let $C \subset S_\mathcal{B}^r$ be a minimal $\left(\frac{\epsilon}{16}\right)$-cover of $S_\mathcal{B}^r \subset \mathcal{L}_\mathcal{B}^r$. Then if $\Omega \in \mathbb{C}^{m \times n}$ satisfies (14) with $S \leftarrow C$ and $\epsilon \leftarrow \frac{\epsilon}{2}$, it will also satisfy*

$$(1 - \epsilon) \|\mathbf{x}\|_2^2 \leqslant \|\Omega \mathbf{x}\|_2^2 \leqslant (1 + \epsilon) \|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathcal{L}_\mathcal{B}^r. \qquad (18)$$

Proof See, e.g., Lemma 3 in [27] for this version. □

Using Lemma 11 one can now easily prove the following corollary of Theorem 9 which demonstrates the existence of random matrices with the OSE property.

Corollary 12 *A random matrix $\Omega \in \mathbb{R}^{m \times n}$ with mean zero, variance m^{-1}, independent sub-gaussian entries has the (ϵ, δ, r)-OSE property for an arbitrary rank r matrix $A \in \mathbb{R}^{n \times N}$ provided that*

$$m \geqslant C \frac{r}{\epsilon^2} \ln \left(\frac{C'}{\epsilon \delta} \right) \geqslant \frac{C}{\epsilon^2} \ln \left(\frac{(\frac{47}{\epsilon})^r}{\delta} \right),$$

where $C' > 0$ is an absolute constant, and $C > 0$ is an absolute constant that only depends on the sub-gaussian norms/parameters of Ω's entries.

Proof As is done in Lemma 11, suppose S is a minimal $\frac{\epsilon}{16}$-cover of the r dimensional unit sphere in the column span of A. The cardinaltiy of this cover is bounded by $\left(\frac{47}{\epsilon} \right)^r$. Apply Theorem 9 to the finite set S. □

Finally, the following lemma demonstrates that matrices satisfying (16) for A also approximately preserve the Frobenius norm of A. We present its proof here as an illustration of basic notation and techniques.

Lemma 13 *Suppose $\Omega \in \mathbb{R}^{m \times n}$ satisfies (16) for a rank r matrix $A \in \mathbb{R}^{n \times N}$. Then,*

$$\left| \|A\|_F^2 - \|\Omega A\|_F^2 \right| \leqslant \epsilon \|A\|_F^2 .$$

Proof Let $\mathbf{e}_i \in \mathbb{R}^N$ denote the standard basis vector where the i-th entry is 1, and all others are zero (i.e., $\mathbf{e}_i$ is the i-th column of the $N \times N$ identity matrix I_N). Similarly let $\mathbf{b}_i$ denote the i-th column of any given matrix B, and set $\tilde{A} := \Omega A$. By (16), we conclude that for all $i \in [N]$ we have

$$\left| \|A\mathbf{e}_i\|_2^2 - \|\Omega A\mathbf{e}_i\|_2^2 \right| \leqslant \epsilon \|A\mathbf{e}_i\|_2^2 .$$

To establish the desired result, we represent the squared Frobenius norm of a matrix as the sum of the squared ℓ_2-norms of its columns. Doing so, we see that

$$\left| \|A\|_F^2 - \|\Omega A\|_F^2 \right| = \left| \sum_{i=1}^N \left(\|\mathbf{a}_i\|_2^2 - \|\tilde{\mathbf{a}}_i\|_2^2 \right) \right| \leqslant \sum_{i=1}^N \left| \|\mathbf{a}_i\|_2^2 - \|\tilde{\mathbf{a}}_i\|_2^2 \right|$$

$$= \sum_{i=1}^N \left| \|A\mathbf{e}_i\|_2^2 - \|(\Omega A)\mathbf{e}_i\|_2^2 \right| \leqslant \epsilon \sum_{i=1}^N \|A\mathbf{e}_i\|_2^2 = \epsilon \|A\|_F^2 .$$

□

We will now define a property of random matrices related to fast approximate matrix multiplication.

2.1.2 The Approximate Matrix Multiplication (AMM) property

First proposed in [26, Lemma 6] (see also, e.g., [28] for other variants), the approximate matrix multiplication property will be crucial to our analysis below.

Definition 14 ((ϵ, δ)-**AMM property**) Let $\epsilon > 0$ and $\delta \in (0, 1)$. A random matrix $\Omega \in \mathbb{R}^{m \times n}$ satisfies the (ϵ, δ)-Approximate Matrix Multiplication property for two arbitrary matrices $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{n \times q}$ if

$$\|A\Omega^T \Omega B - AB\|_F \leqslant \epsilon \|A\|_F \|B\|_F \tag{19}$$

holds with probability at least $1 - \delta$.

The following lemma can be used to construct random matrices with the AMM property from random matrices with the JL property. A slightly generalized version is proven in Appendix A for the sake of completeness.

Lemma 15 (The JL property provides the AMM property) *Let* $A \in \mathbb{R}^{p \times n}$ *and* $B \in \mathbb{R}^{n \times q}$. *There exists a finite set* $S \subset \mathbb{R}^n$ *with cardinality* $|S| \leqslant 2(p + q)^2$ *(determined entirely by A and B) such that the following holds: If a random matrix* $\Omega \in \mathbb{R}^{m \times n}$ *has the* ($\epsilon/2, \delta, 2(p + q)^2$)-*JL property for S, then* Ω *will also have the* (ϵ, δ)-*AMM property for A and B.*

Proof Combine Lemma 48 with Remark 45. $\qquad \square$

Using Lemma 15 one can now prove the following corollary of Theorem 9 which demonstrates the existence of random matrices with the AMM property for any two fixed matrices.

Corollary 16 *Fix* $A \in \mathbb{R}^{p \times n}$ *and* $B \in \mathbb{R}^{n \times q}$. *A random matrix* $\Omega \in \mathbb{R}^{m \times n}$ *with mean zero, variance* $\frac{1}{m}$, *independent sub-gaussian entries will have the* (ϵ, δ)-*Approximate Matrix Multiplication property for A and B provided that*

$$m \geqslant \frac{C}{\epsilon^2} \ln \left(\frac{2(p + q)^2}{\delta} \right),$$

where $C > 0$ *is an absolute constant that only depends on the sub-gaussian norms/parameters of* Ω's *entries.*

Proof Apply Theorem 9 to the finite set S guaranteed by Lemma 15. $\qquad \square$

We now present a capstone definition for this section that will be useful in our analysis of the general measurement ensembles considered herein.

2.1.3 Projection Cost Preserving (PCP) sketches

The following property first appeared in the form below in [29, Definition 1] (see also, however, [30, Definition 13] for the statement of an equivalent property in a different form that appeared earlier).

Definition 17 (**A** (ϵ, c, r)**-Projection Cost Preserving (PCP) sketch**) Let $\epsilon, c > 0$, and $r \in \mathbb{N}$. A matrix $\tilde{X} \in \mathbb{R}^{n \times m}$ is a (ϵ, c, r)-PCP sketch of $X \in \mathbb{R}^{n \times N}$ if for all orthogonal projection matrices $P \in \mathbb{R}^{n \times n}$ with rank at most r,

$$(1 - \epsilon) \|X - PX\|_F^2 \leqslant \left\| \tilde{X} - P\tilde{X} \right\|_F^2 + c \leqslant (1 + \epsilon) \|X - PX\|_F^2 \qquad (20)$$

holds.

The next lemma can be used to construct random matrices that are PCP sketches of a given matrix X with high probability. Before the lemma can be stated, however, we will need one additional definition.

Definition 18 (Head-Tail Split) For any $A \in \mathbb{R}^{m \times n}$, we can split A into his leading r-term and its tail $(n - r)$-term Singular Value Decomposition (SVD) components. That is, consider the SVD of $A = U\Sigma V^T$. For any $r \leqslant \mathrm{rank}(A)$, let $U_r \in \mathbb{R}^{m \times r}$ and $V_r \in \mathbb{R}^{n \times r}$ denote the first r columns of $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$, respectively. We then define $A_r := U_r U_r^T A = A V_r V_r^T$ to be A's best rank r approximation with respect to $\| \cdot \|_F$. Furthermore, we denote the tail term by $A_{\backslash r} := A - A_r$.

One can now see that random matrices with the OSE and AMM properties for matrices derived from $X \in \mathbb{R}^{n \times N}$ will also be PCP sketches of X with high probability. Variants of the following result are proven in [31, 32]. We include the proof in Appendix A.2 for the sake of completeness.

Theorem 19 (Projection-Cost-Preservation via the AMM and OSE properties) *Let $X \in \mathbb{R}^{n \times N}$ of rank $\tilde{r} \leqslant \min\{n, N\}$ have the full SVD $X = U\Sigma V^T$, and let $V_{r'} \in \mathbb{R}^{N \times r'}$ denote the first r' columns of $V \in \mathbb{R}^{N \times N}$ for all $r' \in [N]$. Fix $r \in [n]$ and consider the head-tail split $X = X_r + X_{\backslash r}$. If $\Omega \in \mathbb{R}^{m \times N}$ satisfies*

1. *subspace embedding property (16) with $\epsilon \leftarrow \frac{\epsilon}{3}$ for $A \leftarrow X_r^T$,*
2. *approximate multiplication property (19) with $\epsilon \leftarrow \frac{\epsilon}{6\sqrt{\min\{r,\tilde{r}\}}}$ for $A \leftarrow X_{\backslash r}$ and $B \leftarrow V_{\min\{r,\tilde{r}\}}$,*
3. *JL property (14) with $\epsilon \leftarrow \frac{\epsilon}{6}$ for $S \leftarrow \{$the n columns of $X_{\backslash r}^T\}$, and*
4. *approximate multiplication property (19) with $\epsilon \leftarrow \frac{\epsilon}{6\sqrt{r}}$ for $A \leftarrow X_{\backslash r}$ and $B \leftarrow X_{\backslash r}^T$,*

then $\tilde{X} := X\Omega^T$ is an $(\epsilon, 0, r)$-PCP sketch of X.

Proof See Appendix A.2. □

The following lemma can be used to construct PCP sketches from random matrices with the JL property.

Lemma 20 (The JL property provides PCP sketches) *Let $X \in \mathbb{R}^{n \times N}$ have rank $\tilde{r} \leqslant \min\{n, N\}$. Fix $r \in [n]$. There exist finite sets $S_1, S_2 \subset \mathbb{R}^N$ (determined entirely by X) with cardinalities $|S_1| \leqslant \left(\frac{141}{\epsilon}\right)^{\min\{r,\tilde{r}\}}$ and $|S_2| \leqslant 16n^2 + n$ such that the following holds: If a random matrix $\Omega \in \mathbb{R}^{m \times N}$ has both the $\left(\frac{\epsilon}{6}, \frac{\delta}{2}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property for S_1*

and the $\left(\frac{\epsilon}{6\sqrt{r}}, \frac{\delta}{2}, 16n^2 + n\right)$*-JL property for* S_2*, then* $X\Omega^T$ *will be an* $(\epsilon, 0, r)$*-PCP sketch of* X *with probability at least* $1 - \delta$.

Proof To ensure property 1 of Theorem 19 we can appeal to Lemma 11 to see that Ω being an $(\epsilon/6)$-JL-embedding of a minimal $\left(\frac{\epsilon}{48}\right)$-cover of the at most $\min\{r, \tilde{r}\}$-dimensional unit ball in the column space of X_r^T will suffice. Letting S_1 be this aforementioned $\left(\frac{\epsilon}{48}\right)$-cover, we can further see that $|S_1| \leqslant (141/\epsilon)^{\min\{r, \tilde{r}\}}$ by the proof of Corollary 12. Hence, if $\Omega \in \mathbb{R}^{m \times N}$ has the $\left(\frac{\epsilon}{6}, \frac{\delta}{2}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property for S_1, then property 1 of Theorem 19 will be satisfied with with probability at least $1 - \frac{\delta}{2}$.

Applying Lemma 15 one can see that there exist sets $S_2', S_2'' \subset \mathbb{R}^N$ with $|S_2'| \leqslant 2(n + \min\{r, \tilde{r}\})^2 \leqslant 8n^2$ and $|S_2''| \leqslant 2(n + n)^2 = 8n^2$ such that an $(\epsilon/6\sqrt{r})$-JL-embedding of $S_2' \cup S_2''$ will satisfy both properties 2 and 4 of Theorem 19. Hence, since $r \geqslant 1$, we can see that an $(\epsilon/6\sqrt{r})$-JL-embedding of $S_2 := S_2' \cup S_2'' \cup S$ will satisfy Theorem 19 properties 2 – 4, where S is defined as per property 3. Noting that $|S_2| \leqslant |S_2'| + |S_2''| + |S| \leqslant 16n^2 + n$, we can now see that Ω will satisfy all of Theorem 19's properties 2 – 4 with probability at least $1 - \frac{\delta}{2}$ if it has the $\left(\frac{\epsilon}{6\sqrt{r}}, \frac{\delta}{2}, 16n^2 + n\right)$-JL property for S_2.

Concluding, the prior two paragraphs in combination with the union bound imply that all of Theorem 19's properties 1 – 4 will hold with probability at least $1 - \delta$ if Ω has both the $\left(\frac{\epsilon}{6}, \frac{\delta}{2}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property for S_1 and the $\left(\frac{\epsilon}{6\sqrt{r}}, \frac{\delta}{2}, 16n^2 + n\right)$-JL property for S_2. An application of Theorem 19 now finishes the proof. □

Using Lemma 20 one can now prove the following corollary of Theorem 9 which demonstrates the existence of a PCP sketch for any fixed matrix X.

Corollary 21 *Fix* $X \in \mathbb{R}^{n \times N}$ *and* $r \in [n]$. *Let* $\Omega \in \mathbb{R}^{m \times N}$ *be a random matrix with mean zero, variance* $\frac{1}{m}$, *independent sub-gaussian entries. Then,* $X\Omega^T$ *will be an* $(\epsilon, 0, r)$*-PCP sketch of* X *with probability at least* $1 - \delta$ *provided that*

$$m \geqslant C\frac{r}{\epsilon^2} \max\left\{\ln\left(\frac{C_1}{\epsilon\delta}\right), \ln\left(\frac{C_2 n}{\delta}\right)\right\},$$

where $C_1, C_2 > 0$ *are absolute constants, and* $C > 0$ *is an absolute constant that only depends on the sub-gaussian norms/parameters of* Ω*'s entries.*

Proof Apply Theorem 9 to the finite set S_1 guaranteed by Lemma 20 with $\epsilon \leftarrow \frac{\epsilon}{6}$ and $\delta \leftarrow \frac{\delta}{2}$. Similarly, apply Theorem 9 to the finite set S_2 guaranteed by Lemma 20 with $\epsilon \leftarrow \frac{\epsilon}{6\sqrt{r}}$ and $\delta \leftarrow \frac{\delta}{2}$. The result now follows by Lemma 20. □

We finish here by noting that Corollary 21 is just one example of a PCP sketching result that one can prove with relative ease using Lemma 20. Indeed, Lemma 20 can be combined with other standard results concerning more structured matrices with the JL property (see, e.g., [33–36]) to produce similar theorems where Ω has a fast matrix-vector multiply.

3 The proofs of our main results

The objective is to show that we can retrieve a accurate low rank Tucker approximation of a tensor $\mathcal{X}$ via Algorithm 1 from valid sets of linear leave-one-out and core measurements as described in Section 1.4. We will denote the approximation of $\mathcal{X}$ output by Algorithm 1 as $\mathcal{X}_1$ here to emphasize that a single pass over the original data tensor $\mathcal{X}$ suffices in order to compute the linear input measurements required by Algorithm 1. Hence, Algorithm 1 in this setting is an example of a *streaming algorithm* which doesn't need to store a copy the original uncompressed tensor $\mathcal{X}$ in memory in order to successfully approximate it. Nonetheless, we wish to show that this algorithm still produces a quasi-optimal approximation of $\mathcal{X}$ in the sense of (5) with high probability when given such highly compressed linear input measurements. Particular choices for measurement ensembles will make explicit the dependence on other parameters of the problem (these choices define specializations of Algorithm 1, and are summarized as Algorithm 9 and 11 in Appendix B).

More specifically, in this section we will show that with high probability for a given d-mode tensor $\mathcal{X}$, error tolerance $\epsilon > 0$, and chosen rank truncation parameter r, that

$$\|\mathcal{X} - \mathcal{X}_1\|_2 \leqslant (1 + e^\epsilon) \sqrt{\frac{1 + \epsilon}{1 - \epsilon} \sum_{j=1}^{d} \Delta_{r,j}} \qquad (21)$$

will hold whenever Algorithm 1 is provided with sufficiently informative input measurements. Here the $\Delta_{r,j}$ are defined as per Lemma 3, and "sufficiently informative" means that (i) the leave-one-out measurements used to form $\mathcal{X}_1$ are of sufficient size to satisfy several PCP properties, and that (ii) the core measurements used to form $\mathcal{X}_1$ are of sufficient size to ensure the accurate solution of least squares problems computed as part of Algorithm 1. Finally, we note that one can see from (21) together with Lemma 3 that Algorithm 1 will perfectly recover exactly low Tucker-rank tensors if the rank parameter r is made sufficiently large.

In order to prove that (21) holds, we will need to also consider a weaker variant of Algorithm 1 which permits a *second* pass of the data tensor $\mathcal{X}$. These weaker algorithms will first compute estimates of the factors of the tensor Q_i as Algorithm 1 does, but thereafter will be allowed to use those factors to operate on the original tensor $\mathcal{X}$ in order to approximate its core (see Algorithms 10 and 12). We denote the estimate of the tensor that results from this procedure as $\mathcal{X}_2$ to emphasize that it requires a second accesses to the original data during core recovery. Note that such two-pass algorithms are of less practical value in the big data and compressive sensing settings since it is often not possible to directly access the data tensor again after the initial compressed measurements have been taken in these scenarios. Nevertheless, this two-pass estimate will be extremely useful when proving (21). In particular, our proof will result from the following triangle inequality:

$$\|\mathcal{X} - \mathcal{X}_1\|_2 = \|\mathcal{X} - \mathcal{X}_1 + \mathcal{X}_2 - \mathcal{X}_2\|_2 \leqslant \underbrace{\|\mathcal{X} - \mathcal{X}_2\|_2}_{\text{Term I}} + \underbrace{\|\mathcal{X}_1 - \mathcal{X}_2\|_2}_{\text{Term II}}. \qquad (22)$$

Bounding Term I in (22) will be the subject of Section 3.1. As we shall see, bounding Term I is straightforward if we have that leave-one-out measurements imply various PCP properties; and so the main work of Sections 3.2 and 3.3 will be to demonstrate how, for the structured choices of measurement operators considered herein, we can ensure that the PCP property is satisfied. Bounding Term II, on the other hand, will require us to apply a bound on the error incurred by solving sketched least square problems on a carefully partitioned re-expression of the Term II error. That argument is the subject of Section 3.4. Finally, we combine our analysis of these two error terms along with particular choices for measurement operators to state the full versions of our main results in Section 3.5.

3.1 Bounding $\|\mathcal{X} - \mathcal{X}_2\|_2$

In the two pass scenario, we first compute estimates for the factor matrices, Q_i (see Algorithm 5), using leave-one-out measurements B_i for each $i \in [d]$. Then, using these factor matrices, we wish to solve

$$\arg\min_{\mathcal{H}} \|\mathcal{X} - [\![\mathcal{H}, Q_1, Q_2, \ldots, Q_d]\!]\|_2 .$$

One can see that the solution will be

$$\mathcal{G} := \mathcal{X} \times_1 Q_1^T \times_2 Q_2^T \cdots \times_d Q_d^T$$

(see, e.g., [37]). Let

$$\mathcal{X}_2 := \mathcal{G} \times_1 Q_1 \times_2 Q_2 \cdots \times_d Q_d \tag{23}$$

denote the estimate obtained from a two-pass recovery procedure (i.e., Algorithm 10 or 12). Additionally, we note the following fact about modewise products (see, e.g., [18, Lemma 1]):

$$\mathcal{X} \times_i A \times_i B = \mathcal{X} \times_i (BA).$$

As a result, if we are permitted a second pass over $\mathcal{X}$ to compute the core we have that

$$\begin{aligned}
\mathcal{X}_2 &= \mathcal{G} \times_1 Q_1 \times_2 Q_2 \cdots \times_d Q_d \\
&= \left(\mathcal{X} \times_1 Q_1^T \times_2 Q_2^T \cdots \times_d Q_d^T\right) \times_1 Q_1 \times_2 Q_2 \cdots \times_d Q_d \\
&= \mathcal{X} \times_1 Q_1 Q_1^T \times_2 \cdots \times_d Q_d Q_d^T
\end{aligned}$$

Using this expression we can now bound the two pass error term $\|\mathcal{X} - \mathcal{X}_2\|_2$.

Theorem 22 (Error bound for Two-Pass $\|\mathcal{X} - \mathcal{X}_2\|_2$) *Suppose* $\tilde{X}_{[j]} := X_{[j]}\Omega_{-j}^T \in \mathbb{R}^{n \times m^{d-1}}$ *are* $(\epsilon, 0, r)$-*PCP sketches of* $X_{[j]}$ *for each* $j \in [d]$. *If* $Q_j \in \mathbb{R}^{n \times r}$ *for* $r \leqslant m$ *are factor matrices obtained from Algorithm 5, then*

$$\|\mathcal{X} - \mathcal{X}_2\|_2 = \left\| \mathcal{X} - \mathcal{X} \times_1 Q_1 Q_1^T \times_2 \cdots \times_d Q_d Q_d^T \right\|_2 \leqslant \sqrt{\frac{1+\epsilon}{1-\epsilon} \sum_{j=1}^d \Delta_{r,j}}. \quad (24)$$

Proof Since the $Q_i Q_i^T$ are orthogonal projectors, we have by [38, Theorem 5.1] that

$$\left\| \mathcal{X} - \mathcal{X} \times_1 Q_1 Q_1^T \times_2 \cdots \times_d Q_d Q_d^T \right\|_2^2 \leqslant \sum_{j=1}^d \left\| \mathcal{X} - \mathcal{X} \times_j Q_j Q_j^T \right\|_2^2. \quad (25)$$

From Algorithm 5 (recalling that $\Omega_{(j,j)}^{-1} B_j = X_{[j]}\Omega_{-j}^T = \tilde{X}_{[j]}$), we have that the Q_j's are the best rank-r approximations for their respective sketched problems, since

$$Q_j = \operatorname*{arg\,min}_{\substack{rank(Q)=r \\ Q^T Q = I_r}} \left\| \tilde{X}_{[j]} - Q Q^T \tilde{X}_{[j]} \right\|_F$$

by the Eckart-Young Theorem.

Now suppose that each $U_j \in \mathbb{R}^{n \times r}$ forms an optimal rank r approximation to $X_{[j]}$ in the sense that

$$U_j = \operatorname*{arg\,min}_{\substack{rank(U)=r \\ U^T U = I_r}} \left\| X_{[j]} - U U^T X_{[j]} \right\|_F.$$

By the hypothesis that $\tilde{X}_{[j]}$ is a $(\epsilon, 0, r)$-PCP sketch of $X_{[j]}$, we have that

$$
\begin{aligned}
(1-\epsilon) \left\| \mathcal{X} - \mathcal{X} \times_j Q_j Q_j^T \right\|_2^2 &= (1-\epsilon) \left\| X_{[j]} - Q_j Q_j^T X_{[j]} \right\|_F^2 \\
&\leqslant \left\| \tilde{X}_{[j]} - Q_j Q_j^T \tilde{X}_j \right\|_F^2 \\
&\leqslant \left\| \tilde{X}_{[j]} - U_j U_j^T \tilde{X}_{[j]} \right\|_F^2 \\
&\leqslant (1+\epsilon) \left\| X_{[j]} - U_j U_j^T X_{[j]} \right\|_F^2 = (1+\epsilon)\Delta_{r,j},
\end{aligned}
$$

where we have used the definition of $(\epsilon, 0, r)$-PCP sketches in the first and third inequalities. After a rearrangement of terms, substituting the above into (25) now yields the inequality in (24). □

We have now established in Theorem 22 that we have a quasi-optimal error bound for Term I in (22) whenever our leave-one-out measurement matrices Ω_{-j}^T yield $(\epsilon, 0, r)$-PCP sketches of all d unfoldings X_j. Next, we will demonstrate how to ensure that Kronecker structured and Khatri-Rao structured leave-one-out measurement matrices provide PCP sketches.

3.2 PCP Sketches via Kronecker-structured leave-one-out measurement matrices

In this section we study when Kronecker-structured measurement matrices will provide the PCP property. To begin we will show that the JL and OSE properties are inherited under matrix direct sums and compositions. These are useful facts because our overall leave-one-out matrices can be constructed using these operations. In particular, we will follow the example in the last paragraph of Section 1.4.1 and consider a matrix $\Omega_{-j} \in \mathbb{R}^{m^{d-1} \times n^{d-1}}$ defined as

$$\Omega_{-j} = \bigotimes_{\substack{i=1 \\ i \neq j}}^{d} \Omega_i = \prod_{i'=1}^{d-1} \tilde{\Omega}_{i'} \quad \text{for } \Omega_i \in \mathbb{R}^{m \times n} \tag{26}$$

where

$$\tilde{\Omega}_{i'} := \underbrace{I_n \otimes \cdots \otimes I_n}_{d-1-i'} \otimes \Omega_{i_j(i')} \otimes \underbrace{I_m \otimes \cdots \otimes I_m}_{i'-1} \in \mathbb{R}^{m^{i'} n^{d-1-i'} \times m^{i'-1} n^{d-i'}} \tag{27}$$

for

$$i_j(i') := \begin{cases} i' & \text{if } i' < j \\ i' + 1 & \text{if } i' \geqslant j \end{cases}.$$

Here, I_n denotes an $n \times n$ identity matrix.

The next three lemmas will be used to help show that Ω_{-j} as defined in (26) inherits both the JL and OSE properties from its component Ω_i matrices. Having established this, we can then use, e.g., Lemma 20 to prove PCP sketching results for such Kronecker-structured Ω_{-j}.

Lemma 23 *Suppose that $\Omega_1 \in \mathbb{R}^{m_1 \times N_1}$ and $\Omega_2 \in \mathbb{R}^{m_2 \times N_2}$ are two random matrices. Denote their matrix direct sum by $\Omega = \Omega_1 \oplus \Omega_2 \in \mathbb{R}^{(m_1+m_2) \times (N_1+N_2)}$. Then,*

1. *If Ω_1 and Ω_2 are (ϵ, δ_1, p) and (ϵ, δ_2, p)-JLs respectively, then Ω is an $(\epsilon, \delta_1 + \delta_2, p)$-JL.*
2. *If Ω_1 and Ω_2 are (ϵ, δ_1, r) and (ϵ, δ_2, r)-OSEs respectively, then Ω is an $(\epsilon, \delta_1 + \delta_2, r)$-OSE.*

Proof *Part 1.:* Consider a set $S \subset \mathbb{R}^{N_1+N_2}$ with cardinality p. Let $\mathbf{z} \in S$. Group the first N_1 coordinates of $\mathbf{z}$ into $\mathbf{x} \in \mathbb{R}^{N_1}$ and the last N_2 coordinates of $\mathbf{z}$ into $\mathbf{y} \in \mathbb{R}^{N_2}$. Observe that

$$\|\Omega \mathbf{z}\|_2^2 = \left\| \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\|_2^2 = \|\Omega_1 \mathbf{x}\|_2^2 + \|\Omega_2 \mathbf{y}\|_2^2 \leqslant (1 + \epsilon) \left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 \right) = (1 + \epsilon) \|\mathbf{z}\|_2^2$$

will hold whenever both $\|\Omega_1 \mathbf{x}\|_2^2 \leqslant (1 + \epsilon)\|\mathbf{x}\|_2^2$ and $\|\Omega_2 \mathbf{y}\|_2^2 \leqslant (1 + \epsilon)\|\mathbf{y}\|_2^2$ hold. The $(1 - \epsilon)$-distortion lower bound is similar. As a result, we can use the union bound to see that Ω will have the $(\epsilon, \delta_1 + \delta_2, p)$-JL property.

Part 2.: Suppose $X \in \mathbb{R}^{n \times (N_1 + N_2)}$ has rank r. Let X_1 and X_2 denote the submatrices of X containing the first N_1 and last N_2 columns of X, respectively. Note that both X_1 and X_2 have at most rank r. Furthermore, note also that

$$\left\| \Omega X^T \mathbf{y} \right\|_2^2 = \left\| \begin{pmatrix} \Omega_1 X_1^T \\ \Omega_2 X_2^T \end{pmatrix} \mathbf{y} \right\|_2^2 \leqslant (1 + \epsilon) \left(\left\| X_1^T \mathbf{y} \right\|_2^2 + \left\| X_2^T \mathbf{y} \right\|_2^2 \right) = (1 + \epsilon) \left\| X^T \mathbf{y} \right\|_2^2.$$

will hold for any arbitrary vector $\mathbf{y} \in \mathbb{R}^n$ whenever both $\|\Omega_1 X_1^T \mathbf{y}\|_2^2 \leqslant (1+\epsilon)\|X_1^T \mathbf{y}\|_2^2$ and $\|\Omega_2 X_2^T \mathbf{y}\|_2^2 \leqslant (1+\epsilon)\|X_2^T \mathbf{y}\|_2^2$ hold. The $(1-\epsilon)$-distortion lower bound is similar. As a result, we can see that Ω will be an $(\epsilon, \delta_1 + \delta_2, r)$-OSE by the union bound. □

Note that there is no requirement that Ω_1 and Ω_2 need to be independent in Lemma 23. This is crucial for the next lemma, which will involve many copies of the same measurement matrix.

Lemma 24 (Direct Sums Inherit the OSE and JL Properties) *For some $i' \in [d-1]$, let $\tilde{\Omega}_{i'}$ be defined as in (27) and set $\Omega := \Omega_{i_j(i')} \in \mathbb{R}^{m \times n}$.*

1. *If Ω has the $\left(\epsilon, \frac{\delta}{m^{i'-1} n^{d-i'-1}}, r \right)$-OSE property, then $\tilde{\Omega}_{i'}$ will have the (ϵ, δ, r)-OSE property.*
2. *If Ω has the $\left(\epsilon, \frac{\delta}{m^{i'-1} n^{d-i'-1}}, p \right)$-JL property, then $\tilde{\Omega}_{i'}$ will have the (ϵ, δ, p)-JL, property.*

Proof First consider the rearrangement of $\tilde{\Omega}_{i'}$ in (27) defined as follows

$$\tilde{\Omega} := \underbrace{I_m \otimes I_m \otimes \cdots \otimes I_m}_{i'-1} \otimes \underbrace{I_n \otimes I_n \otimes \cdots \otimes I_n}_{d-1-i'} \otimes \Omega.$$

Note that the Kronecker product of two identity matrices is itself an identity matrix. Thus, we can rewrite this as simply

$$\tilde{\Omega} = \underbrace{I_m \otimes I_m \otimes \cdots \otimes I_m}_{i'-1} \otimes \underbrace{I_n \otimes I_n \otimes \cdots \otimes I_n}_{d-1-i'} \otimes \Omega = \begin{bmatrix} \Omega & 0 & \dots & 0 \\ 0 & \Omega & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \Omega \end{bmatrix}.$$

That is, we have a block diagonal matrix with $\bar{m} = m^{i'-1} n^{d-i'-1}$ copies of Ω along its diagonal. Thus, if Ω has either the $(\epsilon, \delta/\bar{m}, r)$-OSE or the $(\epsilon, \delta/\bar{m}, p)$-JL property, repeated applications of Lemma 23 will then establish the desired OSE or JL property for $\tilde{\Omega}$.

Now consider $\tilde{\Omega}_{i'}$ as in (27). There exist unitary (permutation) matrices L and R which interchange rows and columns such that

$$
L\tilde{\Omega}R = L\left(\underbrace{I_m \otimes I_m \otimes \cdots \otimes I_m}_{i'-1} \otimes \underbrace{I_n \otimes I_n \otimes \cdots \otimes I_n}_{d-1-i'} \otimes \Omega\right)R
$$

$$
= L\left(I_{m^{i'-1}n^{d-1-i'}} \otimes \Omega\right)R = \underbrace{I_n \otimes I_n \otimes \cdots \otimes I_n}_{d-1-i'} \otimes \Omega \otimes \underbrace{I_m \otimes I_m \otimes \cdots \otimes I_m}_{i'-1}
$$

$$
= I_{n^{d-1-i'}} \otimes \Omega \otimes I_{m^{i'-1}} = \tilde{\Omega}_{i'}.
$$

Noting that both the OSE and JL properties are invariant to unitary transformations of a given random matrix, one can now see that $\tilde{\Omega}_{i'} = L\tilde{\Omega}R$ will indeed have the same desired OSE or JL property as was established for $\tilde{\Omega}$. □

Lemma 24 allows us to infer JL and OSE properties of the $\tilde{\Omega}_{i'}$ matrices in (27) from the properties of the smaller random matrices $\Omega_i \in \mathbb{R}^{m \times n}$ appearing in (26). The next lemma will allow us to then use these inferred properties of the $\tilde{\Omega}_{i'}$ matrices to derive OSE and JL properties for Ω_{-j} from (26) in terms of the properties of its component $\Omega_i \in \mathbb{R}^{m \times n}$.

Lemma 25 (A Composition Lemma for the OSE and JL Properties) *Let $\epsilon \in (0, 1)$ and $\tilde{\Omega}_{i'} \in \mathbb{R}^{m^{i'}n^{d-1-i'} \times m^{i'-1}n^{d-i'}}$ for $i' \in [d-1]$.*

1. *If $\tilde{\Omega}_{i'}$ is an $\left(\frac{\epsilon}{2(d-1)}, \frac{\delta}{d-1}, r\right)$-OSE for all $i' \in [d-1]$, then $\tilde{\Omega} = \prod\limits_{i'=1}^{d-1} \tilde{\Omega}_{i'}$ is an (ϵ, δ, r)-OSE.*

2. *If $\tilde{\Omega}_{i'}$ is an $\left(\frac{\epsilon}{2(d-1)}, \frac{\delta}{d-1}, p\right)$-JL for all $i' \in [d-1]$, then $\tilde{\Omega} = \prod\limits_{i'=1}^{d-1} \tilde{\Omega}_{i'}$ is an (ϵ, δ, p)-JL.*

Proof *Part 1.:* Let $Y^T \in \mathbb{R}^{n^{d-1} \times n}$ be an arbitrary matrix of rank at most r. Denote $\tilde{Y}_{i'} = \left(\tilde{\Omega}_{i'} \ldots \tilde{\Omega}_1\right) Y^T \in \mathbb{R}^{m^{i'}n^{d-1-i'} \times n}$ for $i' \in [d-1]$. Note that each $\tilde{Y}_{i'}$ has rank at most r. Fix some $\mathbf{z} \in \mathbb{R}^n$. Suppose for the moment that (16) holds for each $i' \in [d-1]$ with $\Omega = \tilde{\Omega}_{i'}$, $A = \tilde{Y}_{i'-1}$, and $\mathbf{x} = \mathbf{z}$, we have that

$$
\left\|\tilde{\Omega}Y^T\mathbf{z}\right\|_2^2 = \left\|\tilde{\Omega}_{d-1}\left(\tilde{\Omega}_{d-2}\ldots\tilde{\Omega}_1\right)Y^T\mathbf{z}\right\|_2^2
$$

$$
= \left\|\tilde{\Omega}_{d-1}\tilde{Y}_{d-2}\mathbf{z}\right\|_2^2
$$

$$
\leqslant \left(1 + \frac{\epsilon}{2(d-1)}\right)\left\|\tilde{Y}_{d-2}\mathbf{z}\right\|_2^2
$$

$$
= \left(1 + \frac{\epsilon}{2(d-1)}\right)\left\|\tilde{\Omega}_{d-2}\tilde{Y}_{d-3}\mathbf{z}\right\|_2^2
$$

$$
\vdots
$$

$$\leqslant \left(1 + \frac{\epsilon}{2(d-1)}\right)^{d-1} \left\|Y^T \mathbf{z}\right\|_2^2$$

$$\leqslant \left(\frac{1}{1 - \epsilon/2}\right) \left\|Y^T \mathbf{z}\right\|_2^2$$

$$\leqslant (1 + \epsilon) \left\|Y^T \mathbf{z}\right\|_2^2,$$

where we have used the general bound $(1 + k/n)^n \leqslant e^k \leqslant (1-k)^{-1}$ for $k \in [0,1)$ in the second to last inequality. Similarly, for a lower bound one can see that

$$\left\|\tilde{\Omega} Y^T \mathbf{z}\right\|_2^2 \geqslant \left(1 - \frac{\epsilon}{2(d-1)}\right)^{d-1} \left\|Y^T \mathbf{z}\right\|_2^2$$

$$\geqslant (1 - \epsilon) \left\|Y^T \mathbf{z}\right\|_2^2.$$

Union bounding over the failure probability that (16) holds for each $i' \in [d-1]$ as supposed above now yields the desired result.

Part 2.: An essentially identical arguments also applies to obtain the desired JL property result. □

We now have all the necessary results to show how the component maps Ω_i of a Kronecker structured measurement ensemble as per (26) can guarantee a Kronecker sketch with the projection cost preserving property.

Theorem 26 (Kronecker Products of JL matrices yield PCP Sketchs) *Let $\epsilon \in (0,1)$, $X \in \mathbb{R}^{n \times n^{d-1}}$ have rank $r \in [n]$, and $\Omega_{-j} \in \mathbb{R}^{m^{d-1} \times n^{d-1}}$ be defined as in (26) and (27). Furthermore, suppose that the $\Omega_i \in \mathbb{R}^{m \times n}$ in (26) have both the*

1. *$\left(\frac{\epsilon}{12(d-1)}, \frac{\delta}{2(d-1)n^{d-2}}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property, and the*
2. *$\left(\frac{\epsilon}{12\sqrt{r}(d-1)}, \frac{\delta}{2(d-1)n^{d-2}}, 16n^2 + n\right)$-JL property*

for all $i' \in [d-1]$. Then, $X\Omega_{-j}^T$ will be an $(\epsilon, 0, r)$-PCP sketch of X with probability at least $1 - \delta$.

Proof By Lemma 20 we know that $X\Omega_{-j}^T$ will be an $(\epsilon, 0, r)$-PCP sketch of X with probability at least $1 - \delta$ if Ω_{-j} has both the $\left(\frac{\epsilon}{6}, \frac{\delta}{2}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property and the $\left(\frac{\epsilon}{6\sqrt{r}}, \frac{\delta}{2}, 16n^2 + n\right)$-JL property. In fact, by Lemma 25 we can further see that it suffices to have the $\tilde{\Omega}_{i'}$ from (26) and (27) have both the

1. $\left(\frac{\epsilon}{12(d-1)}, \frac{\delta}{2(d-1)}, \left(\frac{141}{\epsilon}\right)^r\right)$-JL property, and the
2. $\left(\frac{\epsilon}{12\sqrt{r}(d-1)}, \frac{\delta}{2(d-1)}, 16n^2 + n\right)$-JL property

for all $i' \in [d-1]$. Finally, looking now at Lemma 24 for each $i' \in [d-1]$ we can see that the assumed properties of the $\Omega_i \in \mathbb{R}^{m \times n}$ in (26) will guarantee both of these sufficient conditions. □

The following corollary of Theorem 26 guarantees a Kronecker sketch with the projection cost preserving property when the component matrices Ω_i in (26) are sub-gaussian random matrices.

Corollary 27 (Kronecker Products of sub-gaussian Matrices Yield PCP Sketches) *Suppose $\mathcal{X}$ is a real valued d-mode tensor with side-lengths all equal to n. Let $\epsilon \in (0, 1), \delta \in (0, 1), r \in [n], j \in [d]$. If $\Omega_{-j} = \bigotimes_{\substack{i=1 \\ i \neq j}}^{d} \Omega_i \in \mathbb{R}^{m^{d-1} \times n^{d-1}}$ defined as in (26) with random matrices $\Omega_i \in \mathbb{R}^{m \times n}$ having i.i.d centered variance m^{-1}, sub-gaussian entries such that*

$$m \geqslant \max \left\{ \frac{C_1 r(d-1)^2}{\epsilon^2} \ln \left(\frac{n^d(d-1)}{\delta} \right), \frac{C_2(d-1)^2}{\epsilon^2} \ln \left(\left(\frac{141}{\epsilon} \right)^r \frac{n^{d-2}(d-1)}{\delta} \right) \right\}$$

for absolute constants $C_1, C_2 > 0$ then the sketched unfolding $\tilde{X}_{[j]} = X_{[j]} \Omega_{-j}^T \in \mathbb{R}^{n \times m^{d-1}}$ is an $(\epsilon, 0, r)$-PCP sketch of $X_{[j]}$ with probability at least $1 - \delta$.

Proof To obtain the first quantity maximized over we apply Theorem 9 with $\epsilon \leftarrow \frac{\epsilon}{12\sqrt{r}(d-1)}, \delta \leftarrow \frac{\delta}{2(d-1)n^{d-2}}$, and $|S| \leftarrow 16n^2 + n$. Similarly, for the second quantity maximized over we apply Theorem 9 with $\epsilon \leftarrow \frac{\epsilon}{12(d-1)}, \delta \leftarrow \frac{\delta}{2(d-1)n^{d-2}}$, and $|S| \leftarrow \left(\frac{141}{\epsilon} \right)^r$. The result now follows from Theorem 26. □

Remark 28 Note that Theorem 26 is quite general, requiring only that the random matrices Ω_i in (26) should be drawn from some distribution having a couple JL properties. As a result of this generality, its Corollary 27 concerning sub-gaussian component matrices turns out to be sub-optimal by (at least a) factor of d in that setting. To obtain a slightly sharper result in d for sub-gaussian Ω_i we recommend replacing our implicit use of Lemma 24 in the proof of Corollary 27 (via Theorem 26) with [39, Lemma 14] instead.

Remark 29 One might wonder how our Corollary 27 based on a Kronecker product construction compares to a dense Gaussian sketch without Kronecker structure. Indeed using dense Gaussian improve the sketching dimension from m^{d-1} to m. The reason is that the sketch we apply to each mode needs to have roughly the same JL-moment guarantees as to that of the final sketch.

However, applying a dense Gaussian sketch has some drawbacks. For example, it naively requires $\Omega(n^d)$ random bits to store the sketching matrix. One could of course try to improve this using, e.g., a pseudorandom number generator. However, each, e.g., matrix entry would then need to be queried every time the sketch was applied, which could incur substantial overhead. In contrast, since none of our sketching matrices require too much memory, we could keep them loaded in uncompressed form on, e.g., a GPU where multiplication with streaming updates could be handled more efficiently.

3.3 PCP sketches via Khatri-Rao structured leave-one-out measurement matrices

In this section we study how to ensure that Khatri-Rao structured leave-one-out measurement matrices will provide the PCP property. To start we will first show that

random Khatri-Rao structured measurement maps, denoted in this section by

$$\Omega_{-j} = \frac{1}{m}\Omega_1 \bullet \Omega_2 \bullet \ldots \Omega_{j-1} \bullet \Omega_{j+1} \bullet \Omega_d \text{ where } \Omega_i \in \mathbb{R}^{m \times n} \ \forall i \in [d] \setminus \{j\},$$

will have the JL property whenever all their component matrices Ω_i have i.i.d. sub-gaussian entries. Having established this, we can then use, e.g., Lemma 20 to prove PCP sketching results for such Khatri-Rao Structured Ω_{-j}.

Theorem 30 *Let $\epsilon > 0, 0 < \delta \leqslant e^{-2}$, and $\Omega_{-j} = \frac{1}{m}\Omega_1 \bullet \Omega_2 \bullet \ldots \Omega_{j-1} \bullet \Omega_{j+1} \bullet \Omega_d$ where all the $\Omega_i \in \mathbb{R}^{m \times n}$ for $i \in [d] \setminus \{j\}$ have i.i.d. mean zero, variance one, sub-gaussian entries. Then Ω_{-j} is an (ϵ, δ, k)-JL whenever*

$$m \geqslant C^{d-1} \max \left\{ \epsilon^{-2} \log \frac{k}{\delta}, \epsilon^{-1} \left(\log \frac{k}{\delta} \right)^{d-1} \right\}$$

for a constant $C \in \mathbb{R}^+$ that depends only on the sub-gaussian norm of the i.i.d. Ω_i-entries.

The proof of Theorem 30 largely follows the argument proposed in Section 2 of [40] concerning the so-called p-moment JL property of Khatri-Rao structured measurements. We note that Kronecker products of sub-gaussian vectors are not sub-gaussian in general, so the general idea is to use Markov's inequality for higher moments of the norm of $\Omega_{-j}\mathbf{y}$ with a fixed $\mathbf{y} \in \mathbb{R}^n$ to obtain the desired result. To proceed with the argument, we will need the following two concentration results.

Lemma 31 *[Lemma 19 in [39]] Let $\mathcal{Y}$ be a $d-1$ mode tensors with side lengths of size n, $p \geqslant 1$, and $\Omega_j(i,:) \in \mathbb{R}^n$ for $j \in [d-1]$, $i \in [m]$ be independent random vectors each satisfying the Khintchine inequality $\left\| \langle \Omega_j(i,:), \mathbf{y} \rangle \right\|_{L^p} \leqslant C_p \left\| \mathbf{y} \right\|_2$ for any vector $\mathbf{y} \in \mathbb{R}^n$ where C_p a constant depending only on p. Then*

$$\| \langle \Omega_1(i,:) \otimes \Omega_2(i,:) \otimes \cdots \otimes \Omega_{d-1}(i,:), vec(\mathcal{Y}) \rangle \|_{L^p} \leqslant C_p^{d-1} \| \mathcal{Y} \|_2 .$$

Lemma 32 *(Corollary 2 in [41]) If $p \geqslant 2$ and $Z, Z_1, \ldots, Z_m$ are i.i.d symmetric random variables then we have*

$$\left\| \sum_{i=1}^m Z_i \right\|_{L^p} \leqslant C \sup_{s \in [\max\{2, \frac{p}{m}\}, p]} \left\{ \frac{p}{s} \left(\frac{m}{p} \right)^{1/s} \| Z \|_{L^s} \right\} .$$

Here $C > 0$ is an absolute constant.

In particular, we will utilize the following corollary of Lemma 32.

Corollary 33 *If under conditions of Lemma 32, in addition, we know that $\| Z \|_{L^s} \leqslant (Cs)^{d-1}$, then*

$$\left\| \frac{1}{m} \sum_{i=1}^m Z_i \right\|_{L^p} \leqslant C^{d-1} \max \left\{ 2^{d-1} \sqrt{\frac{p}{m}}, \frac{(pe)^{d-1}}{m} \right\} .$$

Proof The proof of this corollary loosely follows the argument presented in [40]. Since $\|Z\|_{L^s} \leqslant (Cs)^{d-1}$, $Z_i \sim Z$, the expression over which we are taking the supremum in Lemma 32 is a function whose derivative with respect to s is non-decreasing. That is, the derivative

$$\frac{\partial}{\partial s}\left[\frac{p}{s}\left(\frac{m}{p}\right)^{1/s} s^{d-1}\right] = ps^{d-4}\left(\frac{m}{p}\right)^{1/2}\left((d-2)s - \log\frac{m}{p}\right)$$

has at most a single root in the interval of interest at $s = \frac{\log\frac{m}{p}}{d-2}$. Noting the sign change at this root, we conclude that the maximum value must occur at the endpoints of the interval, and cannot occur at the critical point that is interior to the interval. Evaluating the function of interest at $s = 2$ we obtain $2^{d-2}\sqrt{mp}$; we will further upper-bound this by $2^{d-1}\sqrt{mp}$ in order to simplify analysis for the right endpoint. Additionally, since we are interested only in an upper bound, we need not evaluate the possible endpoint $s = \frac{p}{m}$, since if $2 < \frac{p}{m}$, we are increasing the interval over which we are maximizing by instead considering $s = 2$.

We will now bound the expression at the right endpoint when $s = p \geqslant 2$. Clearly $(1/p)^{1/p} \leqslant 1$ thus

$$\left(\frac{m}{p}\right)^{1/p} p^{d-1} \leqslant m^{1/p} p^{d-1}.$$

If the function value at the right endpoint actually dominates $2^{d-1}\sqrt{mp}$ (i.e., our upper bound of the function value at the left endpoint), we will have that $m^{1/p}p^{d-1} \geqslant 2^{d-1}\sqrt{mp}$ must hold. We will now use this assumption to remove the dependence on m in our current upper bound for the function value at the right endpoint after noting that doing so will still yield a valid upper bound whenever the functions value at the right endpoint fails to already be bounded by $2^{d-1}\sqrt{mp}$.

Proceeding as planned, our assumption yields that

$$m^{1/2-1/p} \leqslant \left(\frac{p}{2}\right)^{d-1} p^{-1/2}.$$

Rearranging of terms, we get that

$$m^{1/p} \leqslant \left[\left(\frac{p}{2}\right)^{d-1} p^{-1/2}\right]^{\frac{2}{p-2}}$$

$$= \left[\left(\frac{p}{2}\right)^{\frac{2d-2}{p-2}}\right] p^{\frac{-1}{p-2}}.$$

The factor $p^{\frac{-1}{p-2}}$ is less than one. A tedious calculation reveals that $\left(\frac{p}{2}\right)^{\frac{2d-2}{p-2}}$ is decreasing for all $p \geqslant 2$, and therefore $m^{1/p}$ bounded by $\lim_{p\to 2}\left(\frac{p}{2}\right)^{\frac{2d-2}{p-2}} = e^{d-1}$. maximizing over our two upper bounds and then averaging over m, we obtain the desired inequality. $\qquad\square$

Now we are ready to give a formal proof of Theorem 30.

Proof of Theorem 30 Note that the result is unchanged for any choice of mode to leave out, so we will shorten the notation and work with $\Omega := \Omega_{-j}$ within this proof. Let K be the sub-gaussian norm of an entry in the Ω_j's. We aim to bound the probability that

$$\left| \frac{1}{m} \|\Omega \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \geqslant \epsilon \|\mathbf{x}\|_2^2 \quad \text{for a fixed } \mathbf{x} \in \mathbb{R}^{n^{d-1}}.$$

Without loss of generality, assume $\|\mathbf{x}\|_2 = 1$. Furthermore, note that

$$\frac{1}{m} \|\Omega \mathbf{x}\|_2^2 - 1 = \frac{1}{m} \sum_{i=1}^{m} \left[\langle \Omega(i,:), \mathbf{x} \rangle^2 - 1 \right]. \tag{28}$$

Now, since the entries of each Ω_j are i.i.d. mean zero and variance one sub-gaussian random variables, they satisfy Khintchine's inequality (see, e.g., [42]) in the form

$$\left\| \langle \Omega_j(i,:), \mathbf{y} \rangle \right\|_{L^p} \leqslant C K \sqrt{p} \, \|\mathbf{y}\|_2 \quad \text{for any fixed } \mathbf{y} \in \mathbb{R}^n.$$

By Lemma 31, this implies that the rows of Ω satisfy a generalized Khintchine's inequality in the form

$$\|\langle \Omega(i,:), \mathbf{x} \rangle\|_{L^p} = \left\| \langle \Omega_1(i,:) \otimes \Omega_2(i,:) \otimes \ldots \otimes \Omega_{d-1}(i,:), \mathbf{x} \rangle \right\|_{L^p} \leqslant (C'p)^{\frac{d-1}{2}} \|\mathbf{x}\|_2, \tag{29}$$

where C' is a new constant that only depends on K.

To bound the L^p-norm of the sum in (28) we will now bound the L^p-norm of each summand. Using the centering Lemma 49 and continuing to estimate for one term we see that

$$\left\| \langle \Omega(i,:), \mathbf{x} \rangle^2 - 1 \right\|_{L^p} \leqslant 2 \left\| \langle \Omega(i,:), \mathbf{x} \rangle^2 \right\|_{L^p} = 2 \|\langle \Omega(i,:), \mathbf{x} \rangle\|_{L^{2p}}^2$$

$$= 2 \|\langle \Omega_1(i,:) \otimes \Omega_2(i,:) \otimes \ldots \otimes \Omega_{d-1}(i,:), \mathbf{x} \rangle\|_{L^{2p}}^2$$

$$\overset{(29)}{\leqslant} 2((C'2p)^{\frac{d-1}{2}})^2 \|\mathbf{x}\|_2^2 \leqslant (C''p)^{d-1} \|\mathbf{x}\|_2^2 = (C''p)^{d-1}.$$

We would now like to apply Corollary 33 to help bound the L^p-norm of the sum in (28). However, we need to symmetrize our random variables first. Toward that end, define

$$Z_i = \rho_i \left(\langle \Omega(i,:), \mathbf{x} \rangle^2 - 1 \right), \tag{30}$$

where ρ_i are i.i.d. Rademacher random variables. Note that $\|Z_i\|_{L^p} = \| \left(\langle \Omega(i,:), \mathbf{x} \rangle^2 - 1 \right) \|_{L^p} \leqslant (C''p)^{d-1}$.

Appealing now to Corollary 33 we have that,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} Z_i \right\|_{L^p} \leqslant (C'')^{d-1} \max \left\{ 2^{d-1} \sqrt{\frac{p}{m}}, \frac{(pe)^{d-1}}{m} \right\} \tag{31}$$

for any $p \geqslant 2$. The upper bound in [43, Lemma 6.3] then further implies that

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \left[\langle \Omega(i, :), \mathbf{x} \rangle^2 - 1 \right] \right\|_{L^p} \leqslant 2 \left\| \frac{1}{m} \sum_{i=1}^{m} Z_i \right\|_{L^p} \leqslant 2(C'')^{d-1} \max \left\{ 2^{d-1} \sqrt{\frac{p}{m}}, \frac{(pe)^{d-1}}{m} \right\}.$$

Employing Markov's inequality, we finally have that

$$\mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \left[\langle \Omega(i, :), \mathbf{x} \rangle^2 - 1 \right] \right| \geqslant \epsilon \right\} = \mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \left[\langle \Omega(i, :), \mathbf{x} \rangle^2 - 1 \right] \right|^p \geqslant \epsilon^p \right\}$$

$$\leqslant (C''')^{p(d-1)} \frac{\max \left\{ \left(\frac{p}{m} \right)^{p/2}, \frac{(pe)^{p(d-1)}}{m^p} \right\}}{\epsilon^p}.$$

Taking $p = \log(k/\delta)$ and $m \geqslant \tilde{C}^{d-1} \max \left\{ \epsilon^{-2} \log \frac{k}{\delta}, \epsilon^{-1} \left(\log \frac{k}{\delta} \right)^{d-1} \right\}$, the last expression is upper bounded by δ/k. Hence, Ω is an (ϵ, δ, k)-JL by the union bound over k vectors. □

Remark 34 In order to employ Lemmas 31 and 32, it is necessary that the *rows* $\Omega_j(i, :)$ have independent and identical distributions and that these rows satisfy Khintchine's inequality. Assuming that the matrices Ω_i all have i.i.d sub-gaussian *entries* as we have done in Theorem 30 implies both these necessary properties of the rows. However, we note that more general (though perhaps less natural) assumptions will also suffice. For example, the distributions of the i.i.d sub-gaussian entries of the Ω_i may also vary by column.

We can now use Theorem 30 to derive row bounds that guarantee that our Khatri-Rao structured sub-gaussian measurement matrices will provide PCP sketches with high probability.

Theorem 35 *Let $\epsilon > 0, 0 < \delta \leqslant e^{-2}$, and $\mathcal{X}$ be a d mode tensor with side-lengths equal to n. Furthermore, suppose that $\Omega_{-j} := \frac{1}{m} \Omega_1 \cdot \Omega_2 \cdot \ldots \Omega_{j-1} \cdot \Omega_{j+1} \cdot \ldots \cdot \Omega_d$ where the $\Omega_i \in \mathbb{R}^{m \times n}$ for $i \in [d] \setminus \{j\}$ are as in Theorem 30 with*

$$m \geqslant C^{d-1} \max \left\{ \epsilon^{-2} \log \frac{\left(\frac{141}{\epsilon} \right)^r}{\delta/2}, \ \epsilon^{-1} \left(\log \frac{\left(\frac{141}{\epsilon} \right)^r}{\delta/2} \right)^{d-1}, \ \frac{r}{\epsilon^2} \log \frac{17n^2}{\delta/2}, \ \frac{r}{\epsilon} \left(\log \frac{17n^2}{\delta/2} \right)^{d-1} \right\}$$

(32)

for a positive constant $C \in \mathbb{R}^+$. Then, $\tilde{X}_{[j]} = X_{[j]} \Omega_{-j}^T$ will be an $(\epsilon, 0, r)$-PCP sketch of $X_{[j]}$ with probability at least $1 - \delta$.

Proof By Lemma 20 we know that it suffices for the measurement matrix Ω_{-j} to have both the $\left(\frac{\epsilon}{6}, \frac{\delta}{2}, \left(\frac{141}{\epsilon} \right)^r \right)$-JL property and $\left(\frac{\epsilon}{6\sqrt{r}}, \frac{\delta}{2}, 16n^2 + n \right)$-JL property. Combining these two required JL properties with Theorem 30 yields (32) after combining and simplifying constants. □

We can now see that both Kronecker and Khatri-Rao structured measurements can satisfy the PCP property required by Theorem 22. This demonstrates that such memory-efficient measurements may be used to bound the first term in (22) as desired. Given this partial success, we will now turn our attention to the second term in (22).

3.4 Bounding $\|\mathcal{X}_1 - \mathcal{X}_2\|_2$

Next, we show how to bound Term II in (22). Recall that $\mathcal{X}_1$ is the output of Algorithm 1, and that $\mathcal{X}_2$ is the tensor recovered by the two-pass algorithm consisting of the first "Factor matrix recovery" phase of Algorithm 1 followed by the second pass core recovery procedure discussed in Section 3.1. That is, $\mathcal{X}_1 := [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ is the single-pass estimate of the tensor $\mathcal{X}$ output by Algorithm 1, and

$$\mathcal{X}_2 = \mathcal{G} \times_1 Q_1 \times_2 \cdots \times_d Q_d = \mathcal{X} \times_1 Q_1 Q_1^T \times_2 \cdots \times_d Q_d Q_d^T,$$

where $\mathcal{G}$ is the core estimate from the two-pass algorithm, and where the $Q_i \in \mathbb{R}^{n \times r}$ have r orthonormal columns.

To begin, we note that the one-pass core $\mathcal{H}$ computed by Algorithm 1 can be recovered from its input measurements by

$$\begin{aligned}
\mathcal{H} &= \mathcal{B}_c \times_1 (\Phi_1 Q_1)^\dagger \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger \\
&= (\mathcal{X} \times_1 \Phi_1 \times_2 \cdots \times_d \Phi_d) \times_1 (\Phi_1 Q_1)^\dagger \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger.
\end{aligned} \quad (33)$$

In addition, we note that the norm of the difference between the two estimates is the same as the norm of the difference of their cores since factor matrices have orthonormal columns.

Lemma 36 *In the notation outlined above,*

$$\|\mathcal{X}_1 - \mathcal{X}_2\|_2 = \|\mathcal{H} - \mathcal{G}\|_2.$$

Proof

$$\begin{aligned}
\|\mathcal{X}_1 - \mathcal{X}_2\|_2 &= \|\mathcal{H} \times_1 Q_1 \times_2 \cdots \times_d Q_d - \mathcal{G} \times_1 Q_1 \times_2 \cdots \times_d Q_d\|_2 \\
&= \|(\mathcal{H} - \mathcal{G}) \times_1 Q_1 \times_2 \cdots \times_d Q_d\|_2 \\
&= \|(Q_1 \otimes \cdots \otimes Q_d)\mathrm{vec}\,(\mathcal{H} - \mathcal{G})\|_2 \\
&= \|\mathcal{H} - \mathcal{G}\|_2
\end{aligned}$$

since $(Q_1 \otimes \cdots \otimes Q_d)$ has orthonormal columns. $\square$

In order to simplify the presentation of our culminating results we next state a definition for an Affine-Embedding property. In Lemma 40 below we then describe how this property relates to the OSE and AMM properties.

Definition 37 ((ϵ, r, N)-**AE property**) Let $\epsilon > 0$, and $r \in \mathbb{N}$, fix $Q \in \mathbb{R}^{n \times r}$ an arbitrary matrix with orthonormal columns and an arbitrary matrix $B \in \mathbb{R}^{n \times N}$. A matrix $\Phi \in \mathbb{R}^{m \times n}$ is an (ϵ, r, N)-Affine Embedding (AE) for given matrices Q and B if it satisfies

$$\left\|(\Phi Q)^\dagger \Phi B\right\|_F \leqslant (1 + \epsilon) \|B\|_F. \tag{34}$$

With this definition in hand, we are now able to prove the main theorem of this section. It will allow us to relate Term II of (22) with Term I.

Theorem 38 *Let* $\mathcal{X}_2 = [\![\mathcal{G}, Q_1, \ldots, Q_d]\!]$ *denote the two-pass tensor estimate, and* $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ *denote the single-pass tensor estimate for a d-mode tensor* $\mathcal{X}$. *Furthermore, let* $\epsilon \in (0, 1)$, *and* $\Phi_i \in \mathbb{R}^{m_c \times n}$ *be a* ($\epsilon/d, r, n^{i-1}r^{d-i}$)-*AE for the matrices* Q_i *and* $\left(X_{[i]} - (X_2)_{[i]}\right) \bigotimes_{j=1}^{i-1} I_n \bigotimes_{j=i+1}^{d} \left(\Phi_j Q_j\right)^\dagger \Phi_j$ *for each* $i \in [d]$. *Then,*

$$\|\mathcal{X}_1 - \mathcal{X}_2\|_2 \leqslant e^\epsilon \|\mathcal{X} - \mathcal{X}_2\|_2. \tag{35}$$

Proof By Lemma 36 it is enough to estimate the difference between the cores $\mathcal{G}$ and $\mathcal{H}$. He have that

$$
\begin{aligned}
\mathcal{H} - \mathcal{G} &\overset{(33)}{=} (\mathcal{X} \times_1 \Phi_1 \times_2 \cdots \times_d \Phi_d) \times_1 (\Phi_1 Q_1)^\dagger \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger - \mathcal{G} \\
&= ((\mathcal{X} - \mathcal{X}_2) \times_1 \Phi_1 \times_2 \cdots \times_d \Phi_d) \times_1 (\Phi_1 Q_1)^\dagger \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger \\
&\quad + (\mathcal{X}_2 \times_1 \Phi_1 \times_2 \cdots \times_d \Phi_d) \times_1 (\Phi_1 Q_1)^\dagger \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger - \mathcal{G} \\
&= (\mathcal{X} - \mathcal{X}_2) \times_1 (\Phi_1 Q_1)^\dagger \Phi_1 \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger \\
&\quad + \mathcal{G} \times_1 (\Phi_1 Q_1)^\dagger \Phi_1 Q_1 \cdots \times (\Phi_d Q_d)^\dagger \Phi_d Q_d - \mathcal{G} \\
&= (\mathcal{X} - \mathcal{X}_2) \times_1 (\Phi_1 Q_1)^\dagger \Phi_1 \times_2 \cdots \times_d (\Phi_d Q_d)^\dagger \Phi_d.
\end{aligned}
$$

Now consider the following related mode-i unfolding for $i \in [d]$, where $\left(\Phi_j Q_j\right)^\dagger \Phi_j$ for $j < i$ is replaced with an $n \times n$ identity matrix I_n,

$$X_i' := \left(X_{[i]} - (X_2)_{[i]}\right) \bigotimes_{j=1}^{i-1} I_n \bigotimes_{j=i+1}^{d} \left(\Phi_j Q_j\right)^\dagger \Phi_j. \tag{36}$$

Each Φ_i is an ($\epsilon/d, r, n^{i-1}r^{d-i}$)-AE, where $Q \leftarrow Q_i$ and $B \leftarrow X_i'$, for $i = 1, 2, \ldots, d$. Thus,

$$
\begin{aligned}
\|\mathcal{H} - \mathcal{G}\|_2 &= \left\|(\Phi_1 Q_1)^\dagger \Phi_1 \left(X_{[1]} - (X_2)_{[1]}\right) \bigotimes_{j=2}^{d} \left(\Phi_j Q_j\right)^\dagger \Phi_j\right\|_F \\
&\leqslant \left(1 + \frac{\epsilon}{d}\right) \|X_1'\|_F \\
&\quad \vdots
\end{aligned}
$$

$$\leqslant \left(1 + \frac{\epsilon}{d}\right)^d \|X_d'\|_F$$
$$\leqslant e^\epsilon \|\mathcal{X} - \mathcal{X}_2\|_2,$$

where we have used Definition 37 d-times together with the bound $\left(1 + \frac{\epsilon}{d}\right)^d \leqslant e^\epsilon$. $\square$

3.5 Putting it all together with row bounds

In the previous subsections we have demonstrated that we can bound both error terms in (22) when the leave-one-out and core measurements satisfy certain embedding properties. In particular, we have shown how the Johnson-Lindenstrauss property (Definition 8) can be used to obtain both the Oblivious Subspace Embedding (Definition 10) and an Approximate Matrix Multiplication properties (Definition 14). These two properties are then used with compositions and direct sums to show how a tensor unfolding will satisfy a Projection Cost Preserving property (Definition 17), which was the essential ingredient in bounding the error term $\|\mathcal{X} - \mathcal{X}_2\|_2$. Next, we introduced Affine-Embeddings (Definition 37), which are the crucial ingredient needed for bounding the error term $\|\mathcal{X}_1 - \mathcal{X}_2\|_2$. In this section we will show how JL and OSE properties imply the AE property. All together then, these will enable us to verify the requirements of Theorem 39 in a straightforward manner once we have specified the type of leave-one-out measurements, and the particular type of sensing matrices.

Theorem 39 (Error bound for one-pass) *Let $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ denote the single-pass tensor estimate for a d-mode tensor $\mathcal{X}$ with side length n obtained from Algorithm 1 using leave-one-out linear measurements B_i for $i \in [d]$, and core measurements $\mathcal{B}_c$. Let $\epsilon > 0$ and $\delta \in (0, 1/2)$. Then,*

$$\|\mathcal{X}_1 - \mathcal{X}\|_2 \leqslant (1 + e^\epsilon) \sqrt{\frac{1+\epsilon}{1-\epsilon} \sum_{j=1}^{d} \Delta_{r,j}} \quad \text{where } \Delta_{r,j} := \sum_{i=r+1}^{\tilde{n}_j} \sigma_i \left(X_{[j]}\right)^2 \quad (37)$$

will hold whenever

1. *$\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ are full-rank matrices for all $i \in [d]$,*
2. *$\Omega_{(i,i)}^{-1} B_i = X_{[i]} \Omega_{-j}^T \in \mathbb{R}^{n \times m^{d-1}}$ are $(\epsilon, 0, r)$-PCP sketches of $X_{[i]}$ for each $i \in [d]$, and*
3. *$\Phi_i \in \mathbb{R}^{m_c \times n}$ is an $(\epsilon/d, r, n^{i-1} r^{d-i})$-AE for the matrices Q_i and X_i' as in (36) for all $i \in [d]$.*

Proof Recalling (22) we have that

$$\|\mathcal{X}_1 - \mathcal{X}\|_2 = \|\mathcal{X}_1 - \mathcal{X} + \mathcal{X}_2 - \mathcal{X}_2\|_2 \leqslant \underbrace{\|\mathcal{X} - \mathcal{X}_2\|_2}_{\text{Term I}} + \underbrace{\|\mathcal{X}_1 - \mathcal{X}_2\|_2}_{\text{Term II}}. \quad (38)$$

We know from Theorem 22 that

$$\|\mathcal{X} - \mathcal{X}_2\|_2 \leqslant \sqrt{\frac{1 + \epsilon}{1 - \epsilon} \sum_{j=1}^{d} \Delta_{r,j}} \tag{39}$$

whenever $X_{[j]} \Omega_{-j}^T \in \mathbb{R}^{n \times m^{d-1}}$ are $(\epsilon, 0, r)$-PCP sketches of $X_{[j]}$ for each $j \in [d]$. Furthermore, Theorem 38 together with our third assumption implies that

$$\|\mathcal{X}_1 - \mathcal{X}_2\|_2 \leqslant e^{\epsilon} \|\mathcal{X} - \mathcal{X}_2\|_2. \tag{40}$$

Using (39) and (40) in (22) now yields the desired result. □

We now have need for the lemma that links the AE property to the JL and OSE properties, and thus provides the necessary machinery to fully account for row bounds that guarantee with high probability that the recovered tensor satisfies the stated bound in Theorem 39.

Lemma 40 *Let $B \in \mathbb{R}^{n \times N}$ and suppose that $Q \in \mathbb{R}^{n \times r}$, $n \geqslant r$, has orthonormal columns. If $\Phi \in \mathbb{R}^{m \times n}$ is an $\left(\frac{1}{2}, \delta, r\right)$-OSE for Q, and has the $\left(\frac{\epsilon}{2\sqrt{r}}, \delta\right)$-AMM property for Q^T and $(I - QQ^T)B$, then Φ will be an (ϵ, r, N)-AE for the matrices Q and B with probability at least $1 - 2\delta$.*

Proof Denote $\tilde{Y} := (\Phi Q)^{\dagger} \Phi B$ and $Y' := Q^T B$, the solutions to the sketched and un-skechted linear least square problems given by minimizing $\|\Phi B - \Phi QY\|_F$ and $\|B - QY\|_F$, respectively, with respect to Y. Whenever Φ is a $(1/2, \delta, r)$-OSE for Q we know that ΦQ will be full-rank. It follows that $(\Phi Q)^T \Phi(B - Q\tilde{Y}) = 0$ will then also hold because $(\Phi Q)^{\dagger} = \left[(\Phi Q)^T (\Phi Q)\right]^{-1} (\Phi Q)^T$. As a consequence,

$$\begin{aligned}
Q^T \Phi^T \Phi Q(\tilde{Y} - Y') &= Q^T \Phi^T \Phi Q(\tilde{Y} - Y') + (\Phi Q)^T \Phi(B - Q\tilde{Y}) \\
&= Q^T \Phi^T \Phi(Q\tilde{Y} - QY' + B - Q\tilde{Y}) \\
&= Q^T \Phi^T \Phi(B - QY').
\end{aligned}$$

When the approximate matrix multiplication property also holds we will now have that

$$\begin{aligned}
\left\|(\Phi Q)^T \Phi Q(\tilde{Y} - Y')\right\|_F = \left\|Q^T \Phi^T \Phi(B - QY')\right\|_F &= \left\|Q^T \Phi^T \Phi(I - QQ^T)B\right\|_F \\
&\leqslant \frac{\epsilon}{2\sqrt{r}} \left\|Q^T\right\|_F \left\|(I - QQ^T)B\right\|_F \\
&= \frac{\epsilon}{2} \left\|(I - QQ^T)B\right\|_F = \frac{\epsilon}{2} \left\|B - QY'\right\|_F.
\end{aligned} \tag{41}$$

Furthermore, whenever Φ is a $\left(\frac{1}{2}, \delta, r\right)$-OSE for the column space of Q, all the eigenvalues of $Q^T \Phi^T \Phi Q - I$ will be within the interval $[-1/2, 1/2]$. Thus, we can

bound its operator norm $\left\| Q^T \Phi^T \Phi Q - I \right\| \leqslant 1/2$. We may now combine this operator norm bound with (41) and see that

$$
\begin{aligned}
\left\| \tilde{Y} - Y' \right\|_F &= \left\| (\tilde{Y} - Y') - (\Phi Q)^T \Phi Q (\tilde{Y} - Y') + (\Phi Q)^T \Phi Q (\tilde{Y} - Y') \right\|_F \\
&\leqslant \left\| (\Phi Q)^T \Phi Q (\tilde{Y} - Y') \right\|_F + \left\| \left[(\Phi Q)^T \Phi Q - I \right] (\tilde{Y} - Y') \right\|_F \\
&\leqslant \frac{\epsilon}{2} \left\| (I - Q Q^T) B \right\|_F + \left\| \left[(\Phi Q)^T \Phi Q - I \right] \right\| \left\| \tilde{Y} - Y' \right\|_F \\
&\leqslant \frac{\epsilon}{2} \left\| (I - Q Q^T) B \right\|_F + \frac{1}{2} \left\| \tilde{Y} - Y' \right\|_F .
\end{aligned}
$$

Rearranging the inequality above while noting the invariance of the Frobenius norm to multiplication by a matrix with orthogonal columns, we learn that

$$
\left\| \tilde{Y} - Y' \right\|_F = \left\| Q(\tilde{Y} - Y') \right\|_F \leqslant \epsilon \left\| (I - Q Q^T) B \right\|_F .
$$

To finish, we may now apply the triangle inequality to see that

$$
\begin{aligned}
\left\| \tilde{Y} \right\|_F &\leqslant \left\| \tilde{Y} - Y' \right\|_F + \left\| Y' \right\|_F \\
&\leqslant \epsilon \left\| (I - Q Q^T) B \right\|_F + \left\| Q^T B \right\|_F \\
&\leqslant \epsilon \left\| (I - Q Q^T) \right\| \| B \|_F + \left\| Q^T \right\| \| B \|_F \\
&= (1 + \epsilon) \| B \|_F .
\end{aligned}
$$

In addition, we note that taking a union bound over the two necessary OSE and AMM conditions establishes the stated probability guarantee. □

We are now prepared to state how a particular choice of distribution used to generate our measurement matrices as well as the leave-one-out measurement type (Kronecker or Khatri-Rao) can satisfy the error bound (37) with high probability. Note that below Algorithms 9 and 11 refer to the specialization of Algorithm 1 to the type of leave-one-out measurement (Kronecker or Khatri-Rao, respectively).

Theorem 41 (Error bound for one-pass Kronecker-structured sub-gaussian measurements) *Suppose $\mathcal{X}$ is a d-mode tensor with side length n. Let $\epsilon > 0$, $\delta \in (0, \frac{1}{3})$, $r \in [n]$. Furthermore, let*

1. $\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ *be arbitrary full-rank matrices,*
2. $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ *for $i \neq j$ be random matrices with mutually independent, mean zero, variance m^{-1}, sub-gaussian entries with*

$$
m \geqslant \max \left\{ \frac{C_1 r (d-1)^2}{\epsilon^2} \ln \left(\frac{n^d d^2}{\delta} \right), \frac{C_2 (d-1)^2}{\epsilon^2} \ln \left(\left(\frac{141}{\epsilon} \right)^r \frac{n^{d-2} d^2}{\delta} \right) \right\}, and
$$

3. $\Phi_i \in \mathbb{R}^{m_c \times n}$ be random matrices with mutially independent, mean zero, variance m^{-1}, sub-gaussian entries with

$$m_c \geqslant \max \left\{ C_3 \ln \left(\frac{(94)^r d}{\delta} \right), \frac{C_4 r d^2}{\epsilon^2} \ln \left(\frac{d(r + n^{d-1})^2}{\delta} \right) \right\}. \qquad (42)$$

Then $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ the output of Algorithm 9 (i.e., Algorithm 1 specialized to Kronecker sub-gaussian measurements $\mathcal{B}_i$ and $\mathcal{B}_c$), will satisfy (37) with probability at least $1 - 3\delta$.

Proof We verify that the requirements of Theorem 39 are satisfied. Note that:

1. $\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ are full-rank matrices by assumption.
2. We have from Corollary 27 where $\delta \leftarrow \frac{\delta}{d}$ that $X_{[j]} \Omega_{-j}^T$ is a $(\epsilon, 0, r)$-PCP sketch of $X_{[j]}$ for all $j \in [d]$ with probability at least $1 - \delta$ when $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ are independent sub-gaussian random matrices with

$$m \geqslant \max \left\{ \frac{C_1 r (d-1)^2}{\epsilon^2} \ln \left(\frac{n^d d^2}{\delta} \right), \frac{C_2 (d-1)^2}{\epsilon^2} \ln \left(\left(\frac{141}{\epsilon} \right)^r \frac{n^{d-2} d^2}{\delta} \right) \right\}.$$

3. A substitution of $\epsilon \leftarrow \frac{1}{2}, \delta \leftarrow \frac{\delta}{d}$ into Corollary 12 yields,

$$m_c \geqslant C_3 \ln \left(\frac{(94)^r d}{\delta} \right) \qquad (43)$$

in order to ensure Φ_i is $(\frac{1}{2}, \frac{\delta}{d}, r)$-OSE. Using Corollary 16, where $\epsilon \leftarrow \frac{\epsilon}{2d\sqrt{r}}$, $\delta \leftarrow \frac{\delta}{d}$ and noting that the matrices Q_i and X_i' as in (36) have are $r \times n$ and $n \times n^{i-1} r^{d-i}$, respectively, we have that when

$$m_c \geqslant \frac{C_4 r d^2}{\epsilon^2} \ln \left(\frac{2d(r + n^{d-1})^2}{\delta} \right) \qquad (44)$$

then Φ_i has the $(\frac{\epsilon}{2d\sqrt{r}}, \frac{\delta}{d})$-AMM property for each $i \in [d]$. Lemma 40 now shows how the OSE and AMM properties ensure that Φ_i has the desired AE property for each $i \in [d]$ with probability at least $1 - 2\delta/d$. The union bound now implies that the third requirement of Theorem 39 will hold with probability at least $1 - 2\delta$.

Taking a maximum over (43) and (44) after simplifying and adjusting constants then yields (42). A final union bound over the failure probabilities for requirements of $\Omega_{(i,j)}$ and Φ_i now yields the result. $\qquad \square$

The following runtime analysis demonstrates that instances of Algorithm 1 can indeed recover low-rank approximations of d-mode tensors of side length n in $o(n^d)$-time. As a result, one can see that Algorithm 1 is effectively a sub-linear time recovery algorithm for a large class of low Tucker-rank tensors.

Theorem 42 *Suppose that $m < n < m^{d-1}$ and that $m > Cm_c$ for some absolute constant $C > 0$. Then, Algorithm 9, when given the B_i and $\mathcal{B}_c$ measurements as input, runs in $\mathcal{O}(dn^2m^{d-1}) - time$ and requires the storage of $dnm^{d-1} + m_c^d$ measurement tensor entries, and at most $d(d-1)mn + n^2 + dm_cn$ total measurement matrix entries.*

Proof Inside the factor matrix recovery loop of Algorithm 5, called by Algorithm 9, the two main sub-tasks are to solve the linear system $\Omega_{(i,i)}F_i = B_i$ and to compute a truncated SVD, $F_i = U\Sigma V^T$. Solving the linear system can be accomplished using QR-factorization via Householder orthogonalization. Doing so requires $\frac{4}{3}n^3$ floating point operations to compute the factorization of $\Omega_{(i,i)}$, and $2n^2m^{d-1}$ operations to form Q^TB_i and n^2m^{d-1} operations to solve $RF_i = Q^TB_i$ via back substitution. The complexity of computing the SVD is $\mathcal{O}(nm^{d-1}\min(n, m^{d-1}))$. Therefore the factor recovery loop has overall complexity

$$\mathcal{O}(dn^2m^{d-1})$$

if we assume that $n < m^{d-1}$.

Next we consider the core recovery loop. for the first iteration of the loop of Algorithm 8, we must form Φ_1Q_1 at a cost of $\mathcal{O}(m_cnr)$. Next we solve a linear system $\Phi_1Q_1H = B_1$ at a cost of $\mathcal{O}(2m_cr^2 - \frac{2}{3}r^3 + 3m_c^dr)$. The first iteration dominates the complexity, since subsequent solves use a smaller right hand side formed from solutions from the previous iterations. Furthermore, if we assume $m_c = \mathcal{O}(m)$, we have a core recovery loop with complexity

$$\mathcal{O}(dm^dr).$$

Thus overall the recovery algorithm has $\mathcal{O}(dn^2m^{d-1})$ complexity. In the situation where $\Omega_{(i,i)} = I_n$ then the computation of the SVDs in the factor recovery step dominates the run-time of the algorithm. Clearly the size of the measurement tensors are nm^{d-1} per factor and m_c^d for the core, which yields the space complexity of the measurements. $\qquad\square$

One of the advantages of the structure of the argument in Theorem 39 is that once it is known how to ensure a given random matrix will satisfy the JL property, we can (with the help of Lemma 40) account for how to assemble related measurement operators that satisfy the error bound (37) in a straightforward way. For example, using Theorem 3.1 in [33] along with bounds appearing in that work on the sketching dimension, we have that a sub-sampled and scrambled Fourier matrix is a (ϵ, δ, p)-JL of vectors in $\mathbb{R}^n$ provided that

$$m \geqslant \frac{C}{\epsilon^2}\log\left(\frac{p}{\delta}\right)\log^4 n.$$

Using such existing results one can easily update Theorem 41 to instead use sub-sampled and scrambled Fourier measurement matrices $\Omega_{(i,j)}$ and Φ_i instead of matrices with independent sub-gaussian entries.

Furthermore, it need not be the case that the distribution is the same for each component map $\Omega_{(i,j)}$ or Φ_i. It is important only that each map satisfies the JL primitive for arbitrary sets. We are free to choose a measurement map for a particular mode to suit some other purpose. For example, in the case that the side lengths of the tensor are unequal, we may prefer to choose a map that admits a fast matrix-vector multiply in order to economize run-time for the modes which are long, and on smaller modes, choose maps which have better trade-offs for quality of approximation in terms of m (e.g., we may prefer dense sub-gaussian random matrices for these modes).

On the other hand, a measurement ensemble which does not rely on the Kronecker product of components, like that in Theorem 30, does not admit this sort of reasoning. Mixing measurement maps of different kinds in that case has no clear advantage, and indeed, may even serve to undermine the advantages.

Theorem 39 can still be used to provide Khatri-Rao structured leave-one-out measurement results. Verifying the requirements of Theorem 39 for Khatri-Rao measurements using Theorem 35 yields the following result.

Theorem 43 (Error bound for one-pass Khatri-Rao structured sub-gaussian measurements) *Suppose $\mathcal{X}$ is a d-mode tensor with side length n, Let $\epsilon > 0$, $\delta \in (0, \frac{1}{3})$, $r \in [n]$. Furthermore, let*

1. *$\Omega_{(i,i)} \in \mathbb{R}^{n \times n}$ be full-rank matrices of any kind,*
2. *$\Omega_{(i,j)} \in \mathbb{R}^{\tilde{m} \times n}$ for $i \neq j$ be random matrices with mutually independent, mean zero, variance $\tilde{m}^{-1}$, sub-gaussian entries with*

$$\tilde{m} \geqslant C^{d-1} \max\left\{ \epsilon^{-2} \ln\left(\frac{\left(\frac{141}{\epsilon} \right)^r 2d}{\delta} \right), \ \epsilon^{-1} \left(\ln\left(\frac{\left(\frac{141}{\epsilon} \right)^r 2d}{\delta} \right) \right)^{d-1}, \right.$$
$$\left. \frac{r}{\epsilon^2} \ln\left(\frac{34n^2 d}{\delta} \right), \ \frac{r}{\epsilon} \left(\ln\left(\frac{34n^2 d}{\delta} \right) \right)^{d-1} \right\},$$

   and
3. *$\Phi_i \in \mathbb{R}^{m_c \times n}$ be random matrices with mutually independent, mean zero, variance m^{-1}, sub-gaussian entries with*

$$m_c \geqslant \max\left\{ C_3 \ln\left(\frac{(94)^r d}{\delta} \right), \ \frac{C_4 r d^2}{\epsilon^2} \ln\left(\frac{d(r + n^{d-1})^2}{\delta} \right) \right\}.$$

Then $\mathcal{X}_1 = [\![\mathcal{H}, Q_1, \ldots, Q_d]\!]$ the output of Algorithm 11 (i.e., Algorithm 1 specialized to Khatri-Rao sub-gaussian leave-one-out measurements $\mathcal{B}_i$ and Kronecker measurements $\mathcal{B}_c$) will satisfy (37) with probability at least $1 - 3\delta$.

Proof The proof is again based on verifying the requirements of Theorem 39.

1. The first requirement is satisfied by assumption.
2. The row requirement for the $\Omega_{(i,j)}$ follows from an application of Theorem 35 with $\delta \leftarrow \delta/d$. As a result of doing so, we learn that the second requirement will be satisfied with probability at least $1 - \delta$ after a union bound.

3. The third requirement is satisfied identically to the argument in the proof of Theorem 41.

   The proof now concludes identically to the proof of Theorem 41. □

Comparing the Theorem 42, we note, e.g., that Theorem 43 will require the storage of $d\tilde{m}n + m_c^d$ measurement tensor entries (where the $\tilde{m}$ here is from the second condition of Theorem 43). Comparing this to the $dnm^{d-1} + m_c^d$ measurements from Theorem 42 (where m here is as in Theorem 41) one can see that there are parameter regimes where Khatri-Rao structured measurements will lead to a smaller overall measurement budget.

4 Numerical experiments

In this section we present numerical results that support our theoretical contributions and address practical trade-offs involved in the different choices for measurement type. (Code and resulting data are available at https://github.com/cahaselby/leave_one_out_recovery.) Unless otherwise specified, the tensors in the experiments are random three-mode cubic tensors, with side length $n = 300$ and rank $r = 10$. We use the following procedure (same as in [16]) to generate low-rank tensors; the core's entries are uniformly and independently drawn from $[0, 1]$ and the factors are formed by first sampling a standard normal distribution for each entry and then normalized and made orthogonal using QR-factorization. The data points in the plots are the mean of 100 independent trials. The parameter m refers to the sketching dimension for maps $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ used in recovering factor Q_i. For the left-out mode we remove the need to solve the full $n \times n$ linear system by setting $\Omega_{(i,i)}$ to be the $n \times n$ identity matrix. The parameter m_c refers to the sketching dimension for $\Phi_i \in \mathbb{R}^{m_c \times n}$ which are used in recovering the core $\mathcal{H}$.

In experiments with noise, the additive Gaussian noise tensor $\mathcal{N}$ is scaled according to the desired signal to noise ratio and added to the true, (low-rank) tensor $\mathcal{X}_0$. That is, $\mathcal{X} = \mathcal{X}_0 + \mathcal{N}$ is the observed, noisy tensor.

Signal to noise ratio (SNR) is calculated as $10 \log_{10} (\|\mathcal{X}\|_2 / \|\mathcal{X}_0 - \mathcal{X}\|_2)$. Relative error is calculated as $(\|\hat{\mathcal{X}} - \mathcal{X}\|_2)/ \|\mathcal{X}_0\|_2$ where $\hat{\mathcal{X}}$ is the full estimated tensor.

4.1 Recovering low-rank tensors

In this first simple experiment, we fix the signal to noise ratio at 30 decibels (dB) and vary the sketching dimension m to show the dependence on the accuracy of our estimate on the number of measurements. For each m we set $m_c = 2m$. Rank truncation is fixed at $r = 10$, which matches the rank of the true, noiseless tensor $\mathcal{X}_0$, see Fig. 2. For the plot (b) in the figure, we have the maximum principal angle among the three estimated factors and true factors (Q_i, U_i) in degrees, see [44]. Note, there is no straightforward way to plot the relative error which is due to the factor estimates vs. the core estimate, because the decomposition will in general not be unique. However, since principal angle is invariant to non-singular transformations, plot (b) provides empirical evidence that the factor estimates alone are improving with
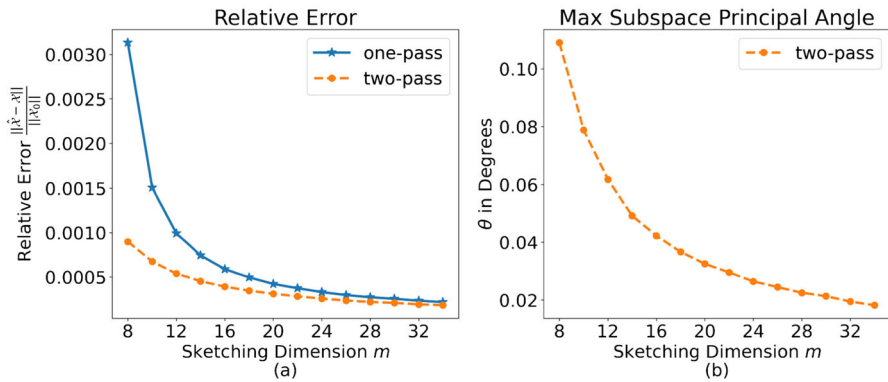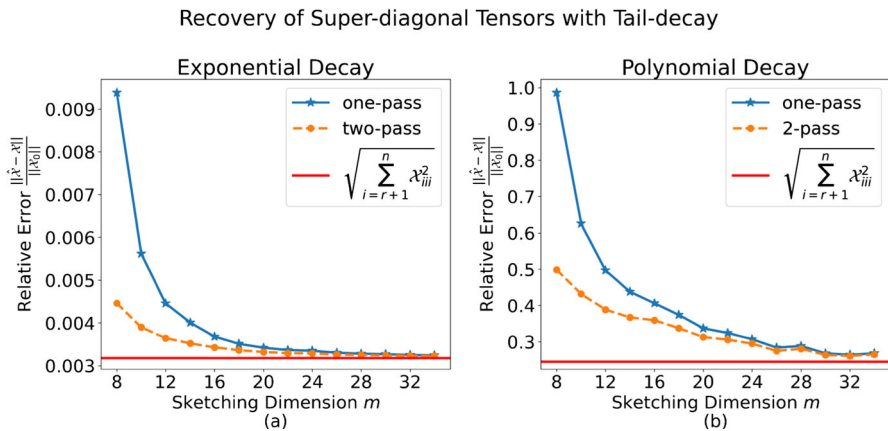
Noisy Recovery of Low Rank Tensors, SNR = 30 dB

Fig. 2 Error plots for different sketching dimensions and a fixed SNR of 30 dB and a fixed rank truncation of 10. Plot (a) compares relative errors for both one-pass and two-pass recovery. Plot (b) shows the maximum principal angle among all estimated sub-spaces and the true factor matrices

sketching dimension. We note that for these low-rank tensors with noise, we are able to fit at or below the level of noise (relative error of 0.001) easily — evidently finding good rank 10 approximations to the (full-rank) noisy tensor $\mathcal{X}$.

This perhaps surprising result motivated us to try the method on a class of tensors in which we could be more certain about what quality of rank 10 approximation is achievable. In our second set of experiments, we examine performance on super-diagonal tensors with tail decay. Since we are truncating to rank 10, this tail can be thought of as structured, deterministic noise. These are tensors where all values are zero except for those on the diagonal, and where we have some decay on the magnitude of all values on the diagonal for indices larger than $r = 10$. In particular we have two types of decay, exponential tail decay in plot (a), where

$$
\mathcal{X}_{ijk} = \begin{cases} 1 & i = j = k \in [r+1] \\ 10^{-1(i-r)} & i = j = k \in [r+2, n] \\ 0 & \text{otherwise} \end{cases} \tag{45}
$$

and polynomial tail decay in plot (b),

$$
\mathcal{X}_{ijk} = \begin{cases} 1 & i = j = k \in [r+1] \\ (i-r)^{-1} & i = j = k \in [r+2, n] \\ 0 & \text{otherwise.} \end{cases} \tag{46}
$$

These highly constrained tensors are clearly not low-rank, however it is reasonable to suppose that a recovery algorithm for a given rank truncation would output an estimate that is close to the leading r terms of the diagonal. The residual in that case will simply be the norm of the tail-sum $\sqrt{\sum_{i=r+1}^{n} \mathcal{X}_{iii}^2}$, which we have included as the red horizontal line in Fig. 3.

Fig. 3 Error plots for different sketching dimensions in the noiseless setting with a fixed rank truncation of 10. Plot (a) compares relative errors for both one-pass and two-pass recovery for super-diagonal tensors where the diagonal entries have exponential decay of type (45), and plot (b) super-diagonal tensors with polynomial decay of the type (46)

4.2 Allocating core and factor measurements

One question raised by our error analysis is how to weigh the error contribution between the tasks of estimating the factor matrices and estimating the core. In other words, for a given total measurement budget, how should we allocate between the two tasks if we wish to decrease overall relative error? In the following experiment (see Fig. 4) we find the relative error under various noise levels for pairs of sketching dimensions (m, m_c). We compare pairs (13, 12) and (11, 36) and (8, 48). These choices of sketching dimensions were chosen since they have nearly equal overall compression ratios of 0.57%, 0.58%, 0.62% respectively, however they vary considerably on whether they emphasize measurements to be used in estimating the factors or the core of the tensor. Note that the two-pass error, which relies only on the factor matrix estimates is naturally best when the factor sketches are larger, i.e. the $m = 13$ case. However the relative error of the recovered tensor in the one-pass setting is more than ten times better when more of the total measurement budget is allocated to estimate the core as shown in Fig. 4. This shows that in some situations it is preferable to allocate more resources to obtain measurements for the core than the factors, up to some threshold. For example in Fig. 4, the rank of the true signal is 10, and going below this dimension for the factor sketches does correspond with no longer improving on the accuracy in terms of the trade-off between m and m_c.

4.3 Error bounds apply to sub-gaussian measurement matrices

In this next experiment we demonstrate, in a similar manner as done in Figure 1 in [16], that recovery performance of Algorithm 11 does not vary greatly for different choices of types of sub-gaussian measurement matrices. What is different from that

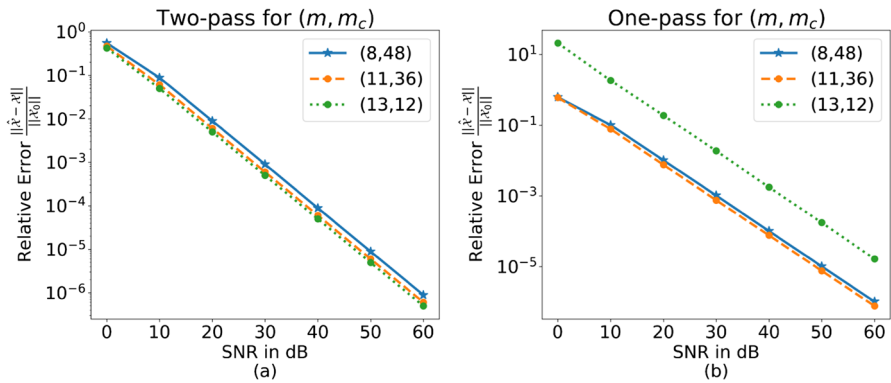## Balancing Measurements for Factor and Core Recovery

Fig. 4 Relative error plots at different signal to noise ratios for two-pass and one-pass recovery of noisy low-rank tensors. Ordered pairs indicates the choice for sketching dimensions (m, m_c)

earlier work is that the measurement ensembles are all Kronecker structured. Plotted in Fig. 5 are relative errors for Gaussian (g), sparse uniform from $[-1, 0, 1]$ with weights $\frac{1}{6}, \frac{2}{3}, \frac{1}{3}$ (sp0), sub-sampled randomized Fourier transforms as in [45] (rfd) and a mixed measurement ensemble that uses Gaussian-RFD-sparse measurements where we vary by mode which measurement type is used, which is a scenario that is practically and theoretically not well suited for the Khatri-Rao structured measurement operators used in [16].

Comparing Measurement Types
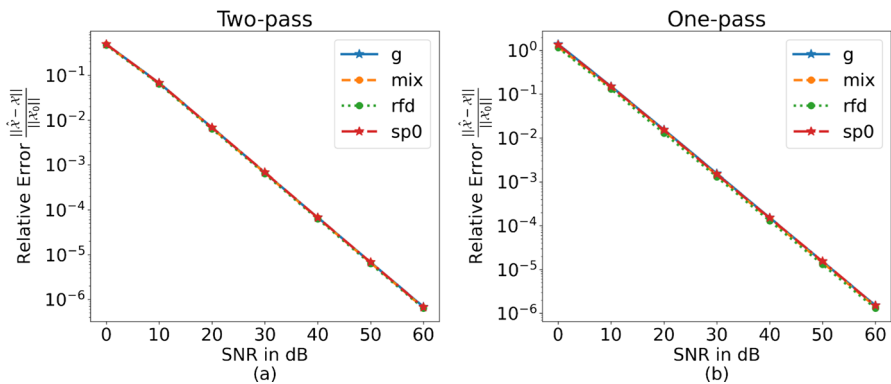
Fig. 5 Relative errors for one-pass, two-pass for Kronecker measurement ensembles made up of different kinds of sub-gaussian random matrices. The legend g, sp0, rfd, mix correspond to Gaussian, sparse, sub-sampled random Fourier transform, and a mixture of all three for the measurement ensembles
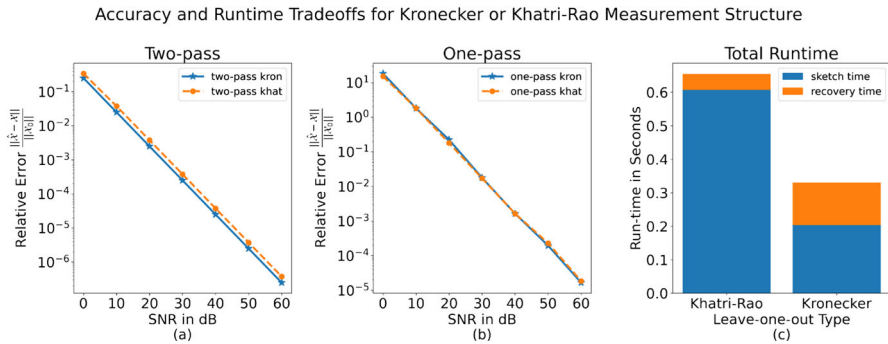
Fig. 6 Relative error comparison for Khatri-Rao and Kronecker-structured measurements. Right hand figure shows the average time for sketching and recovery phases of the algorithm

4.4 Comparison to Khatri-Rao

This set of experiments demonstrates that the sketching phase will dominate the runtime of Algorithm 1 regardless of the choice of leave-one-out type, however the Kronecker-structured measurements are able to generate more measurements for a fixed number of operations as compared to Khatri-Rao structured measurements, see Fig. 6. This means that it is possible to achieve similar or better performance using strictly modewise measurements and in less overall time as problems grow in size with respect to total number of tensor elements; i.e. both number of modes and length of those modes. In Fig. 6 for the Kronecker-structured measurements we sketch to $m = 25$, for the Khatri-Rao ensemble we sketch pairs of modes to 225. We see that the Kronecker measurements perform incrementally better in terms of relative error but at less than half the overall run-time. Sketching times are about five times faster for the Kronecker-structured measurements as compared to the Khatri-Rao. Note that this does trade speed for space — the total number of entries in the leave one out measurements is nearly three times larger for the Kronecker-structured measurements versus the Khatri-Rao, i.e. sketches B_i as per (8) and (11) have sizes 300×25^2, and 300×225 respectively.

4.5 Application to video summary task

As a practical demonstration, we consider the same video summary task first described in [10] and again in [16]. In this demonstration, the video is taken with a camera in a fixed position. The video is a nature scene and a person walks in front of the camera at two different time points in the second half of the video. The first 100 and the last 193 frames are removed since they include setup that results in small shifts of the camera. The entire video has been converted to grayscale. This yields a three mode tensor of dimensions $2200 \times 1080 \times 1980$ which has a size of about 41 GB when stored as an array of doubles. We wish to identify the parts of the scene that include the person walking and distinguish them from the relatively static scene elsewhere. As discussed in [16], there is a third salient time varying feature in this particular video, which is the

light intensity of the scene, since at around frame 940 the scene darkens. Furthermore there are changes in the light intensity as the camera automatically adjusts after the person walks in and out of frame. For this reason, we cluster the frames using three centers, rather than two.

In all cases, we use k-means to cluster the frames, however we assign features to frames in four different ways:

1. Using the sketch $B_1 \in \mathbb{R}^{2200 \times 20^2}$, as in (7) that leaves out the time dimension, then clustering using k-means on the rows of the unfolding of the sketch along the first, temporal mode.
2. Unfolding the temporal mode of the reconstructed tensor using a one-pass set of measurements, i.e. $(X_1)_{[1]} \in \mathbb{R}^{2200 \times 2138400}$ (recall that $\mathcal{X}_1$ denotes the output of Algorithm 1).

(a)

(b)
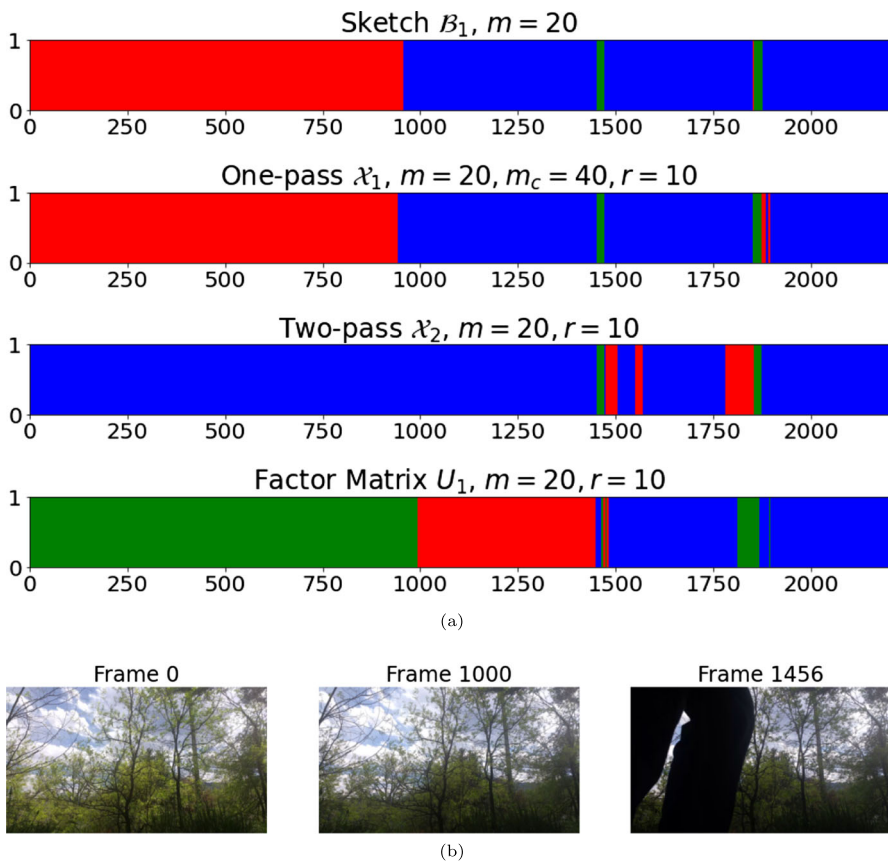
Fig. 7 (a) Cluster assignments for the 2200 frames in the video. Top run corresponds to using the measurements for the first mode only B_1, middle rows use one and two pass approximations of the tensor, and the last row uses the factor matrix U_1 for the temporal mode only. We use Gaussian sketching matrices for both spatial modes and the real part of RFD for the temporal mode. Sketching parameters are $m = 20$, $m_c = 40$ and rank truncation of $r = 10$ in all modes. (b) Three reference frames at 0, 1000 and 1496

3. Unfolding the reconstructed tensor using a two-pass scenario, $(X_2)_{[1]} \in \mathbb{R}^{2200 \times 2138400}$ (recall that $\mathcal{X}_2$ denotes the output of Algorithm 10).
4. Estimated temporal factor matrix $U_1 \in \mathbb{R}^{2200 \times 20}$ (see in Algorithm 1).

As we can see in Figs. 7a, b, the sketch alone shows reliable clustering of the main temporal changes in the video, which verifies the observation in [16] about using the measurements as an effective feature set for clustering, although in that case the measurements were Khatri-Rao structured whereas ours are Kronecker-structured. The unfoldings of the reconstructed tensor also reliably distinguishes the main parts of the scene. The reconstruction is useful at least to get clusted interpretability. Although certainly natural to wish to cluster on the temporal factor, this method appears inferior to any of the preceding.

As an added advantage of using the modewise, Kronecker structured measurements, we can in principle select measurement maps for different modes. Gaussian measurement maps theoretically have some advantages over other types in terms of accuracy for a fixed number of measurements, whereas applying RFD or other Fourier-like transforms to modes that have longer fibers would net a better payoff in terms of overall run-time because of the faster matrix-vector multiply permitted by these structured matrices. In this demonstration, we use Gaussian matrices along the spatial modes, and RFD matrices for the temporal mode.

In the earlier work [10], the authors describe a variant of Tucker-Alternating Least Squares (aka Higher Order Orthogonal Iteration) that employed TensorSketch to produce the necessary measurements used to reconstruct the same video tensor data we have used here. In the subsequent work [16], different authors then perform the same task, but use an approach which fits the framework we have described as Algorithm 1, where the measurement matrices are Khatri-Rao structured, and the Ω_i have entries drawn from standard Gaussian distribution. Here, analysis of the type afforded by Theorem 43 may also help explain the discrepancy between the sketching dimensions seen in [10] and [16]. Naturally there are several differences between the approaches, but the CountSketch matrices used in TensorSketch operators as shown in [39] have a $\mathcal{O}\left(\frac{1}{\delta}\right)$ dependency in order to ensure the OSE property, whereas the other ensembles, such as dense Gaussians, enjoy an $\mathcal{O}\left(\log \frac{1}{\delta}\right)$ dependence for this parameter.

As was discussed in [16], the video is not especially low-rank in practice — in particular along the spatial dimensions in terms of relative error of the reconstruction.
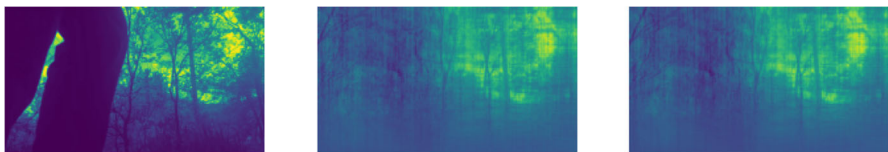

Fig. 8 The 1456th frame of the grey-scale video is shown for the original, the one-pass, and two-pass reconstructions using sketching dimension of $m = 300$, $m_c = 601$ and $r = 50$ for each mode. Although the reconstructions for this choice of sketching dimension and rank truncation are not particularly accurate for this video, nevertheless the reconstructions provide enough information to perform the summary task of clustering the frames into the major temporal changes that occur during the scene

However the clusters appear distinct enough that assigning clusters with this summary type of information is still possible (Fig. 8).

Appendix A: Technical proofs

Herein we provide proofs for selected results from Section 2.1.

A.1: Proof of Lemma 15

Our proof of Lemma 15 will utilize several intermediate lemmas. Our first lemma concerning the approximate preservation of inner products is a slight generalization of [46, Corollary 2].

Lemma 44 (The JL property implies angle preservation) *Let $S \subset \mathbb{C}^n$ with cardinality at most p and $\epsilon \in (0, 1)$. If a random matrix $\Omega \in \mathbb{C}^{m \times n}$ has the $(\epsilon/4, \delta, 4p^2)$-JL property for*

$$S' = \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + i\frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - i\frac{\mathbf{y}}{\|\mathbf{y}\|_2} \mid \mathbf{x}, \mathbf{y} \in S \right\},$$

then

$$|\langle \Omega \mathbf{x}, \Omega \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| \leqslant \epsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \quad \forall \mathbf{x}, \mathbf{y} \in S \tag{A1}$$

will be satisfied with probability at least $1 - \delta$.

Proof Note that if either $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, then (A1) automatically holds because $0 \leqslant 0$. Thus, suppose without loss of generality that $\mathbf{x}, \mathbf{y} \neq 0$. Considering the normalizations $\mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \mathbf{v} = \frac{\mathbf{y}}{\|\mathbf{y}\|_2}$, one can see that the polarization identity implies that

$$
\begin{aligned}
|\langle \Omega \mathbf{u}, \Omega \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle| &= \left| \frac{1}{4} \sum_{\ell=0}^{3} i^{\ell} \left(\|\Omega \mathbf{u} + i^{\ell} \Omega \mathbf{v}\|_2^2 - \|\mathbf{u} + i^{\ell} \mathbf{v}\|_2^2 \right) \right| \\
&\leqslant \frac{1}{4} \sum_{\ell=0}^{3} \frac{\epsilon}{4} (\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2)^2 = \epsilon
\end{aligned}
$$

will hold whenever (14) holds with $S \leftarrow S'$ and $\epsilon \leftarrow \epsilon/4$. The result now follows by renormalizing. □

Remark 45 Note that if $S \subset \mathbb{R}^n$ it suffices for a random matrix $\Omega \in \mathbb{R}^{m \times n}$ to have the $(\epsilon/2, \delta, 2p^2)$-JL property for a smaller set $S' \subset \mathbb{R}^n$ in Lemma 44. This can be seen by using the real version of the polarization identity instead of the complex version.

The next lemma constructs a set S to utilize in Lemma 44 based on two matrices with normalized columns. The end result is an entrywise approximate matrix multiplication property for the two column-normalized matrices in question.

Lemma 46 (The JL property allows approximate matrix multiplies for unitary matrices) *Let $V \in \mathbb{C}^{n \times p}$ and $U \in \mathbb{C}^{n \times q}$ have unit ℓ^2-normalized columns. Suppose that $\Omega \in \mathbb{C}^{m \times n}$ satisfies* (A1) *from Lemma 44 where $S = \{\mathbf{u}_j | \mathbf{u}_j = U[:, j]\} \cup \{\mathbf{v}_j | \mathbf{v}_j = V[:, j]\}$. Then*

$$\left| \left(V^* \Omega^* \Omega U - V^* U\right)_{k,j} \right| \leqslant \epsilon, \ \text{for all } 1 \leqslant k \leqslant p \ \text{and} \ 1 \leqslant j \leqslant q.$$

Proof Note that $|S| = p + q$. Thus, $|S'| \leqslant 4(p+q)^2$ in Lemma 44. Furthermore,

$$\Omega V = \begin{pmatrix} | & | & \cdots & | \\ \Omega\mathbf{v}_1 & \Omega\mathbf{v}_2 & \ldots & \Omega\mathbf{v}_p \\ | & | & \cdots & | \end{pmatrix}, \ \text{and} \ \Omega U = \begin{pmatrix} | & | & \cdots & | \\ \Omega\mathbf{u}_1 & \Omega\mathbf{u}_2 & \ldots & \Omega\mathbf{u}_q \\ | & | & \cdots & | \end{pmatrix}.$$

Hence, $\left((\Omega V)^* \Omega U\right)_{k,j} = \langle \Omega\mathbf{u}_j, \Omega\mathbf{v}_k \rangle$. Therefore, given Lemma 44, for all choices of k, j we have

$$\left| \left(V^* \Omega^* \Omega U - V^* U\right)_{k,j} \right| = \left| \langle \Omega\mathbf{u}_j, \Omega\mathbf{v}_k \rangle - \langle \mathbf{u}_j, \mathbf{v}_k \rangle \right| \leqslant \epsilon \|\mathbf{v}_k\|_2 \|\mathbf{u}_j\|_2 = \epsilon.$$

$\square$

The next lemma constructs a new set S to utilize in Lemma 44 by selecting a well chosen subset of the singular vectors of both A and B. This set will ultimately determine how the finite set S promised by Lemma 15 depends on A and B. As we shall see, it's proven by applying Lemma 46 to two unitary matrices provided by the SVDs of A and B.

Lemma 47 (The JL property implies the AMM property for arbitrary matrices) *Let $A \in \mathbb{C}^{p \times n}$ and $B \in \mathbb{C}^{n \times q}$ have SVDs given by $A = U_1 \Sigma_1 V^*$ and $B = U \Sigma_2 V_2^*$, and suppose that $\Omega \in \mathbb{C}^{m \times n}$ satisfies the conditions of Lemma 46 for U and V. Then,*

$$\|A\Omega^* \Omega B - AB\|_F \leqslant \epsilon \|A\|_F \|B\|_F$$

Proof We will expand the quantity of interest according the SVD of A and B. Doing so we see that

$$
\begin{aligned}
\|A\Omega^* \Omega B - AB\|_F &= \|U_1 \Sigma_1 V^* \Omega^* \Omega U \Sigma_2 V_2^* - U_1 \Sigma_1 V^* U \Sigma_2 V_2^*\|_F \\
&= \|U_1 \Sigma_1 \left(V^* \Omega^* \Omega U - V^* U\right) \Sigma_2 V_2^*\|_F \\
&= \|\Sigma_1 \left(V^* \Omega^* \Omega U - V^* U\right) \Sigma_2\|_F \\
&= \sqrt{\sum_{k=1}^{p} \sum_{j=1}^{q} (\Sigma_1)_{k,k}^2 |V^* \Omega^* \Omega U - V^* U|_{k,j}^2 (\Sigma_2)_{j,j}^2}
\end{aligned}
$$

$$\leqslant \sqrt{\sum_{k=1}^{p} \sum_{j=1}^{q} \sigma_k(A)^2 \epsilon^2 \sigma_j(B)^2}$$

$$= \epsilon \sqrt{\sum_{k=1}^{p} \sigma_k(A)^2} \sqrt{\sum_{j=1}^{q} \sigma_j(B)^2}$$

$$= \epsilon \|A\|_F \|B\|_F.$$

□

Lemmas 44, 46, and 47 now collectively prove the following generalized version of Lemma 15.

Lemma 48 (The JL property provides the AMM property) *Let $A \in \mathbb{C}^{p \times n}$ and $B \in \mathbb{C}^{n \times q}$. There exists a finite set $S \subset \mathbb{C}^n$ with cardinality $|S| \leqslant 4(p + q)^2$ (determined entirely by A and B) such that the following holds: If a random matrix $\Omega \in \mathbb{C}^{m \times n}$ has the $(\epsilon/4, \delta, 4(p + q)^2)$-JL property for S, then Ω will also have the (ϵ, δ)-AMM property for A and B.*

We will make use of this simple centering result with regards to L^p norms of random variables.

Lemma 49 *Suppose X a real random variable, and let $p \geqslant 1$, Then,*

$$\|X - \mathbb{E}[X]\|_{L^p} \leqslant 2 \|X\|_{L^p}$$

Proof Let $\mu = \mathbb{E}[X]$. By Jensen's inequality, $\|\mu\|_{L^p} = |\mu| \leqslant \mathbb{E}[|X|] = \|X\|_{L^1}$ Observe,

$$\|X - \mu\|_{L^p} \leqslant \|X\|_{L^p} + \|\mu\|_{L^p}$$
$$\leqslant \|X\|_{L^p} + \|X\|_{L^1}$$
$$\leqslant 2 \|X\|_{L^p}$$

Where we have used Minkowski's inequality in the first line, and Jensen's inequality in the third. □

A.2: Proof of Theorem 19

A similar proof appears in support of [32, Theorem 2], which was itself simplified from earlier work [31]. We reproduce the proof here for completeness, and to clarify details. We begin by restating Theorem 19 for ease of reference.

Theorem 50 (Restatement of Theorem 19) *Let $X \in \mathbb{R}^{n \times N}$ of rank $\tilde{r} \leqslant \min\{n, N\}$ have the full SVD $X = U \Sigma V^T$, and let $V_{r'} \in \mathbb{R}^{N \times r'}$ denote the first r' columns of $V \in \mathbb{R}^{N \times N}$ for all $r' \in [N]$. Fix $r \in [n]$ and consider the head-tail split $X = X_r + X_{\backslash r}$. If $\Omega \in \mathbb{R}^{m \times N}$ satisfies*

1. *subspace embedding property* (16) *with* $\epsilon \leftarrow \frac{\epsilon}{3}$ *for* $A \leftarrow X_r^T$,
2. *approximate multiplication property* (19) *with* $\epsilon \leftarrow \frac{\epsilon}{6\sqrt{\min\{r,\tilde{r}\}}}$ *for* $A \leftarrow X_{\backslash r}$ *and*
   $B \leftarrow V_{\min\{r,\tilde{r}\}}$,
3. *JL property* (14) *with* $\epsilon \leftarrow \frac{\epsilon}{6}$ *for* $S \leftarrow \left\{ \text{the n columns of } X_{\backslash r}^T \right\}$, *and*
4. *approximate multiplication property* (19) *with* $\epsilon \leftarrow \frac{\epsilon}{6\sqrt{r}}$ *for* $A \leftarrow X_{\backslash r}$ *and* $B \leftarrow$
   $X_{\backslash r}^T$,

then $\tilde{X} := X\Omega^T$ *is an* $(\epsilon, 0, r)$-*PCP sketch of* X.

Proof Let $Q \in \mathbb{R}^{n \times r}$ be a an arbitrary matrix with orthonormal columns so that $QQ^T \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix. It suffices to show that

$$\left| \left\| (I - QQ^T)X \right\|_F^2 - \left\| (I - QQ^T)X\Omega^T \right\|_F^2 \right| \leq \epsilon \left\| (I - QQ^T)X \right\|_F^2. \tag{A2}$$

Writing X in terms of its head-tail split, (A2) becomes

$$\left| \left\| (I - QQ^T)(X_r + X_{\backslash r}) \right\|_F^2 - \left\| (I - QQ^T)(X_r + X_{\backslash r})\Omega^T \right\|_F^2 \right| \leq \epsilon \left\| (I - QQ^T)X \right\|_F^2, \tag{A3}$$

where X_r has the thin SVD representation $X_r = U_r \Sigma_r V_r^T$ with U_r and V_r containing the first r columns of U and V from the full SVD of X, respectively.[4]

Letting $P := (I - QQ^T)$, and using that $\operatorname{tr} AA^T = \|A\|_F^2$, one may expand the left hand side of (A3). Noting that $X_{\backslash r} X_r^T = 0$ and $P = P^T$ while doing so, we can now see that

$$\left| \left\| (I - QQ^T)(X_r + X_{\backslash r}) \right\|_F^2 - \left\| (I - QQ^T)(X_r + X_{\backslash r})\Omega^T \right\|_F^2 \right| =$$
$$\left| \operatorname{tr} \left(P(X_r X_r^T + X_{\backslash r} X_{\backslash r}^T)P \right) - \operatorname{tr} \left(P(X_r \Omega^T \Omega X_r^T + X_r \Omega^T \Omega X_{\backslash r}^T + X_{\backslash r} \Omega^T \Omega X_r^T + X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T)P \right) \right|.$$

Regrouping terms in the last expression while noting the invariance of trace to transposition, one can we can now further see that

$$\left| \operatorname{tr} \left(P(X_r X_r^T + X_{\backslash r} X_{\backslash r}^T)P \right) - \operatorname{tr} \left(P(X_r \Omega^T \Omega X_r^T + X_r \Omega^T \Omega X_{\backslash r}^T + X_{\backslash r} \Omega^T \Omega X_r^T + X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T)P \right) \right|$$
$$= \left| \operatorname{tr} \left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P \right) + 2\operatorname{tr} \left(P(X_{\backslash r} \Omega^T \Omega X_r^T)P \right) + \operatorname{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T)P \right) \right|$$
$$\leq \left| \operatorname{tr} \left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P \right) \right| + 2\left| \operatorname{tr} \left(P(X_{\backslash r} \Omega^T \Omega X_r^T)P \right) \right| + \left| \operatorname{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T)P \right) \right|.$$

---

[4] If $r \geq N$ we let $V_r = V$.

Looking back at (A3) in the light of this last computation, one can now see that it suffices to prove that

$$\left|\operatorname{tr}\left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P\right)\right| + 2\left|\operatorname{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P\right)\right| + \left|\operatorname{tr}\left(P(X_{\backslash r}X_{\backslash r}^T - X_{\backslash r}\Omega^T \Omega X_{\backslash r}^T)P\right)\right|$$
$$\leqslant \epsilon \|PX\|_F^2 \tag{A4}$$

holds to establish the desired result. Going forward we will therefore aim to prove (A4) by proving each of the following three bounds:

(a) $\left|\operatorname{tr}\left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P\right)\right| \leqslant \frac{\epsilon}{3}\|PX\|_F^2$,
(b) $\left|\operatorname{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P\right)\right| \leqslant \frac{\epsilon}{6}\|PX\|_F^2$, and
(c) $\left|\operatorname{tr}\left(P(X_{\backslash r}X_{\backslash r}^T - X_{\backslash r}\Omega^T \Omega X_{\backslash r}^T)P\right)\right| \leqslant \frac{\epsilon}{3}\|PX\|_F^2$.

Proving (a) – (c) will establish (A4), thereby completing the proof.

Proof of Bound (a): Using again that $\operatorname{tr} AA^T = \operatorname{tr} A^T A = \|A\|_F^2$ and that $P = P^T$ we have

$$\left|\operatorname{tr}\left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P\right)\right| = \left|\left\|X_r^T P\right\|_F^2 - \left\|\Omega X_r^T P\right\|_F^2\right|.$$

Applying Lemma 13 in the light of assumption 1, we now have that

$$\left|\operatorname{tr}\left(P(X_r X_r^T - X_r \Omega^T \Omega X_r^T)P\right)\right| \leqslant \frac{\epsilon}{3}\left\|X_r^T P\right\|_F^2 = \frac{\epsilon}{3}\left\|V_r V_r^T X^T P\right\|_F^2$$
$$\leqslant \frac{\epsilon}{3}\left\|X^T P\right\|_F^2 = \frac{\epsilon}{3}\|PX\|_F^2$$

as desired.

Proof of Bound (b): Using the invariance of trace to both transposition and permutations, as well as that $P^T = P = P^2$, we can see that

$$\left|\operatorname{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P\right)\right| = \left|\operatorname{tr}\left(P(X_r \Omega^T \Omega X_{\backslash r}^T)P\right)\right| = \left|\operatorname{tr}\left(P X_r \Omega^T \Omega X_{\backslash r}^T\right)\right|.$$

Recalling the full SVD $X = U\Sigma V^T$, we note that if $\tilde{r} := \operatorname{rank}(X) < \min\{n, N\}$ we can remove the last $n - \tilde{r}$ columns of U, the last $N - \tilde{r}$ columns of V, and the last $n - \tilde{r}$ rows and $N - \tilde{r}$ columns of Σ to form the thin SVD $X = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ with $\tilde{U} \in \mathbb{R}^{n \times \tilde{r}}$, $\tilde{\Sigma} \in \mathbb{R}^{\tilde{r} \times \tilde{r}}$, and $\tilde{V} \in \mathbb{R}^{N \times \tilde{r}}$. Having done so we note that $\tilde{\Sigma}$ will be invertable and that $\tilde{U}\tilde{U}^T X_r = X_r$ so that we may write

$$\left|\operatorname{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P\right)\right| = \left|\operatorname{tr}\left(P X_r \Omega^T \Omega X_{\backslash r}^T\right)\right|$$
$$= \left|\operatorname{tr}\left(P\tilde{U}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T\right)\right|$$
$$= \left|\operatorname{tr}\left(P\tilde{U}\tilde{\Sigma}\tilde{V}^T \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T\right)\right|$$

$$= \left| \mathrm{tr}\left(\left(P\tilde{U}\tilde{\Sigma}\tilde{V}^T \right) \left(\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right) \right) \right|$$

$$= \left| \mathrm{tr}\left((PX) \left(\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right) \right) \right|.$$

Recall now that $\langle A, B \rangle_F := \mathrm{tr}(AB^T)$ is an inner product on matrices with $\|A\|_F = \sqrt{\mathrm{tr}(AA^T)}$. Hence, we may apply the Cauchy-Schwarz inequality to our last expression to see that

$$\left| \mathrm{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P \right) \right| = \left| \mathrm{tr}\left((PX)\left(\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right) \right) \right|$$

$$\leqslant \|PX\|_F \left\| \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right\|_F$$

$$= \|PX\|_F \left\| \tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right\|_F.$$

Expanding X_r in terms of its thin SVD representation $X_r = U_r \Sigma_r V_r^T$ we now have that

$$\left| \mathrm{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P \right) \right| \leqslant \|PX\|_F \left\| \tilde{\Sigma}^{-1}\tilde{U}^T X_r \Omega^T \Omega X_{\backslash r}^T \right\|_F$$

$$= \|PX\|_F \left\| \tilde{\Sigma}^{-1}\tilde{U}^T U_r \Sigma_r V_r^T \Omega^T \Omega X_{\backslash r}^T \right\|_F$$

$$= \|PX\|_F \left\| V_{\min\{r,\tilde{r}\}}^T \Omega^T \Omega X_{\backslash r}^T \right\|_F$$

$$= \|PX\|_F \left\| X_{\backslash r}\Omega^T \Omega V_{\min\{r,\tilde{r}\}} \right\|_F.$$

Finally, we may now use that $X_{\backslash r} V_{\min\{r,\tilde{r}\}} = X(I_N - V_r V_r^T)V_{\min\{r,\tilde{r}\}} = X\left(V_{\min\{r,\tilde{r}\}} - V_{\min\{r,\tilde{r}\}} \right) = 0$ to see that

$$\left| \mathrm{tr}\left(P(X_{\backslash r}\Omega^T \Omega X_r^T)P \right) \right| \leqslant \|PX\|_F \left\| X_{\backslash r}\Omega^T \Omega V_{\min\{r,\tilde{r}\}} \right\|_F$$

$$= \|PX\|_F \left\| X_{\backslash r}\Omega^T \Omega V_{\min\{r,\tilde{r}\}} - X_{\backslash r} V_{\min\{r,\tilde{r}\}} \right\|_F$$

$$\leqslant \|PX\|_F \left(\frac{\epsilon}{6\sqrt{\min\{r,\tilde{r}\}}} \|X_{\backslash r}\|_F \|V_{\min\{r,\tilde{r}\}}\|_F \right),$$

where we have utilized assumption 2 in the last inequality. We are now finished after using that $\|V_{\min\{r,\tilde{r}\}}\|_F = \sqrt{\min\{r,\tilde{r}\}}$, and noting that $\|X_{\backslash r}\|_F \leqslant \|PX\|_F$ holds for all rank $n - r$ or greater orthogonal projections P by the definition of X_r.

Proof of Bound (c): Again using the invariance of trace to permutations as well as $P = P^2 = I - QQ^T$ we have that

$$
\begin{aligned}
\left| \mathrm{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) P \right) \right| &= \left| \mathrm{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) \right) \right| \\
&= \left| \mathrm{tr} \left((I - QQ^T)(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) \right) \right| \\
&\leqslant \left| \mathrm{tr} \left(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T \right) \right| \\
&\quad + \left| \mathrm{tr} \left(QQ^T (X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) \right) \right| \\
&\leqslant \left| \left\| X_{\backslash r}^T \right\|_F^2 - \left\| \Omega X_{\backslash r}^T \right\|_F^2 \right| + \left\| QQ^T \right\|_F \left\| X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T \right\|_F,
\end{aligned}
$$

where we have again utilized Cauchy-Schwarz in the last inequality. Utilizing assumption 3, the first term just above can be bounded by $\frac{\epsilon}{6} \|X_{\backslash r}^T\|_F^2 = \frac{\epsilon}{6} \|X_{\backslash r}\|_F^2$ using an argument analogous to the proof of Lemma 13. Doing so we see that

$$
\begin{aligned}
\left| \mathrm{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) P \right) \right| &\leqslant \frac{\epsilon}{6} \|X_{\backslash r}\|_F^2 + \left\| QQ^T \right\|_F \left\| X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T \right\|_F \\
&= \frac{\epsilon}{6} \|X_{\backslash r}\|_F^2 + \sqrt{r} \left\| X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T \right\|_F.
\end{aligned}
$$

Finally, we may now employ assumption 4 to bound the second term just above. Doing so we obtain that

$$
\begin{aligned}
\left| \mathrm{tr} \left(P(X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T) P \right) \right| &\leqslant \frac{\epsilon}{6} \|X_{\backslash r}\|_F^2 + \sqrt{r} \left\| X_{\backslash r} X_{\backslash r}^T - X_{\backslash r} \Omega^T \Omega X_{\backslash r}^T \right\|_F \\
&\leqslant \frac{\epsilon}{6} \|X_{\backslash r}\|_F^2 + \sqrt{r} \frac{\epsilon}{6\sqrt{r}} \|X_{\backslash r}\|_F^2 \\
&= \frac{\epsilon}{3} \|X_{\backslash r}\|_F^2.
\end{aligned}
$$

To conclude we note again that $\left\| X_{\backslash r} \right\|_F \leqslant \| PX \|_F$ holds for all rank $n - r$ or greater orthogonal projections P by the definition of X_r. $\qquad\square$

Appendix B: Algorithms

There are four tasks relevant in Algorithm 1 where by making difference choices for these tasks, we get variants of the algorithm. Two of the tasks are related to the measurement process: sketching the tensor in order to produce leave-one-out measurements, and also sketching the tensor without leaving out a mode in order to produce measurements useful in computing the core. The second two tasks are then recovering the factor matrices, and recovering the core.

Note, if we permit a second pass on the tensor $\mathcal{X}$ after estimating factor matrices Q_i, then core measurements are not necessary, and the optimal core given these factor

matrices can be found by computing

$$\mathcal{G} = \mathcal{X} \times_1 Q_1^T \times_2 Q_2^T \cdots \times_d Q_d^T$$

See section 4.2 in [1]. This then is the computation used in two-pass versions of the algorithm; it requires first to compute the factor matrices from measurements, and then apply these factors modewise to the original tensor; no separate measurement tensor for the core is required. Crucially however, this relies on a second access to the data; for which we are computing a HOSVD.

In this appendix then, we break apart Algorithm 1 into these tasks, and show how combining different choices for these tasks produces the variants of the algorithm considered.

Within the pseudo-code, "unfold" refers to the operation of taking a tensor and flattening it into a matrix of the size listed by arranging the specified mode's fibers as the columns. The operation "fold" is the inverse of this, taking a matrix as viewed as an unfolding along the specified mode and reshaping it into a tensor of the given dimensions.

Algorithm 2 takes measurement matrices (e.g. sub-gaussian random matrices) and the data tensor $\mathcal{X}$ and produces a set of leave-one-out measurements which can be viewed as tensors or as flattened matrices. That is, it is practical to apply the measurement matrices along the modes and obtain a tensor of measurements with $d - 1$ modes each of length m and one mode of length n, see Fig. 1 for a schematic depiction. The measurement process takes slices of the tensor and maps them to (smaller) slices with (mostly) shorter edge lengths. Or we can conceive of the measurements as a matrix by unfolding the measurement tensor along the mode that is uncompressed and thus obtaining a matrix of size $n \times m^{d-1}$, one such matrix for each mode $i \in [d]$. It is Kronecker structured because of the correspondence of modewise products of tensors with matrices to matrix products of unfoldings of a tensor with matrices, see for example (3).

Alternatively, we can use Algorithm 3 which also produces a set of leave-one-out measurements of the tensor $\mathcal{X}$ which can be viewed as a matrix. It is Khatri-Rao structured because the measurement matrix applied to the unfolding is formed using Khatri-Rao products. Note, unlike Algorithm 2, there is not necessarily any natural way to view the measurements B_i as tensors of d modes - the sketching process in this case takes slices of the tensor and maps them to vectors and we gather these into matrices B_i of size $n \times m$ for each $i \in [d]$.

In order to estimate the core, we wish another, independent set of measurements. These are obtained in the manner as Algorithm 2, only now we are permitted to compress all the modes, producing a tensor of d modes all with side length equal to m_c.

Next, we describe the procedure which takes as input (*a*) leave-one-out measurements of $\mathcal{X}$ (from Algorithm 2 or 3), (*b*) the full-rank sensing matrices applied to the uncompressed modes of $\mathcal{X}$, and (*c*) our desired target rank vector of **r**, and then outputs a factor matrix Q_i for each mode. In the case of exact arithmetic, no rank truncation, and no noise, this exactly recovers the factors (see (24)).

Algorithm 2 Leave-One-Out Kronecker Sketching.

---

input :

$\Omega_{(i,j)}$ where $\text{rank}(\Omega_{(i,i)}) = n$ and $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ for $i, j \in [d], i \neq j$

$\mathcal{X}$ a d mode tensor with side lengths n

output: B_i for $i \in [d]$

```
# COMPLETE MEASUREMENT SUMMARY
```
for $i \in [d]$ **do**
$\quad \mathcal{B}_i \leftarrow \mathcal{X} \times_1 \Omega_{(i,1)} \times_2 \Omega_{(i,2)} \times_3 \cdots \times_d \Omega_{(i,d)}$
```
    # Now unfold the measurement tensor so the mode-i
      fibers are columns, size n × m^{d−1}
```
$\quad B_i \leftarrow \text{unfold}(\mathcal{B}_i, n \times m^{d-1}, \text{mode} = i)$
end

```
# AN EXAMPLE ONE-PASS VARIANT
```
Initialize $B_j \leftarrow$ Matrix of 0s $\forall j \in [d]$
for $(i_1, \ldots, i_d) \in [n]^d$ **do**
$\quad$ **for** $j \in [d]$ **do**
$\quad\quad \mathbf{y} \leftarrow (\Omega_{(j,1)})_{:,i_1} \odot \cdots \odot (\Omega_{j,j-1})_{:,i_{j-1}} \odot (\Omega_{(j,j+1)})_{:,i_{j+1}} \odot \cdots \odot (\Omega_{(j,d)})_{:,i_d}$
$\quad\quad B_j \leftarrow B_j + \mathcal{X}_{i_1,\ldots,i_j,\ldots,i_d} (\Omega_{(j,j)})_{:,i_j} \mathbf{y}^T$
$\quad$ **end**
end

Algorithm 3 Leave-One-Out Khatri-Rao Sketching.

---

input :

$\Omega_{(i,j)}$ where $\text{rank}(\Omega_{(i,i)}) = n$ and $\Omega_{(i,j)} \in \mathbb{R}^{m \times n}$ for $i, j \in [d], i \neq j$

$\mathcal{X}$ a d mode tensor with side lengths n

output: B_i for $i \in [d]$

for $i \in [d]$ **do**
$\quad B_i \leftarrow \Omega_{(i,i)} X_{[i]} \left(\Omega_{(i,1)} \bullet \Omega_{(i,2)} \bullet \ldots \bullet \Omega_{i,i-1} \bullet \Omega_{i,i+1} \bullet \ldots \bullet \Omega_{i,d} \right)^T$
end

Algorithm 4 Core Sketching.

---

input :

$\Phi_i \in \mathbb{R}^{m_c \times n}, i \in [d]$

$\mathcal{X}$ a d mode tensor with side lengths n

output: $\mathcal{B}_c$

$\mathcal{B}_c \leftarrow \mathcal{X} \times_1 \Phi_1 \times_2 \Phi_2 \times_3 \cdots \times_d \Phi_d$

Lastly, we consider the task of obtaining the core of the HOSVD of the data tensor. The two ways described in Section 3 are to either compute the core using a second access to the data tensor - in which case this is a matter of applying the transpose of

---

Algorithm 5 Recover HOSVD Factors from Leave-One-Out Measurements.

> **input** :
> $B_i \in R^{n \times m^{d-1}}$ for $i \in [d]$ measurements that leave mode i uncompressed
> $\Omega_{(i,i)} \in R^{n \times n}$ for $i \in [d]$
> $\mathbf{r} = (r, \ldots, r)$ desired rank for HOSVD
> **output**: $Q_1, \ldots, Q_d \in \mathbb{R}^{n \times r}$
> \# Factor matrix recovery
> **for** $i \in [d]$ **do**
> > \# Solve $n \times n$ linear system
> > Solve $\Omega_{(i,i)} F_i = B_i$ for F_i
> > \# Compute SVD and keep the top r singular vectors
> > $U, \Sigma, V^T \leftarrow \text{SVD}(F_i)$
> > $Q_i \leftarrow U_{:,:r}$
> **end**

---

the factor matrices from Algorithm 5 to the data tensor. This is detailed in Algorithm 10.

Algorithm 6 Compute HOSVD Core with Second Access.

> **input** :
> $Q_1, \ldots, Q_d \in \mathbb{R}^{n \times r}$ computed factor matrices
> $\mathcal{X}$ a d mode tensor with side lengths n.
> **output**: $\mathcal{G}$ a d mode tensor with side lengths r
> $\mathcal{G} = \mathcal{X} \times_1 Q_1^T \times_2 Q_2^T \cdots \times_d Q_d^T$

In the scenario in which a second access to the tensor is not desired, instead we obtain the core of the HOSVD by solving a linear system involving the measurement operators and the factor matrices, see (33). This is equivalent to solving the linear system a mode at a time as detailed in Algorithm 7, a method of practical value because it does not require as much working memory.

A third possibility is to "re-use" leave-one-out measurements to compute the core. Theoretically this involves new dependencies on the errors introduced by estimating the factors and errors introduced by recovering the core that are not addressed in any of our main results. Practically however, it would be desirable to avoid having to produce the core measurement tensor, and empirically on synthetic data the overall error is not effected.

We are now able to state the variants of the general algorithm.

Algorithm 7 Recover HOSVD Core from Measurements.

---

input :
$\mathcal{B}_c$ a d mode tensor with side lengths m_c
$\Phi_i \in \mathbb{R}^{m_c \times n}$
$Q_1, \ldots, Q_d \in \mathbb{R}^{n \times r}$ computed factor matrices
output: $\mathcal{H}$
for $i \in [d]$ **do**

    `# unfold measurements, mode-`i` fibers are columns,`
       `size `$m_c \times r^{(i-1)} m_c^{d-1-(i-1)}$
    $H \leftarrow \text{unfold}(\mathcal{B}_c, m_c \times r^{(i-1)} m_c^{d-1-(i-1)}, \text{mode} = i)$
    `# Undo the mode-`i` measurement operator and factor's`
       `action by finding least square solution to `$m_c \times r$
       `over-determined linear system`
    Solve $\Phi_i Q_i H_{\text{new}} = H$ for H_{new}
    `# reshape the flattened partially solved core into`
       `a tensor`
    $\mathcal{B}_c \leftarrow \text{fold}(H_{\text{new}}, \underbrace{r \times r \cdots \times r}_{i} \times \underbrace{m_c \times \cdots \times m_c}_{d-i}, \text{mode} = i)$
    `# Each iteration `$m_c \rightarrow r$` in `i`th mode`

end

Algorithm 8 Recover HOSVD Core from Recycled Leave One Out Measurements.

---

input :
B_j for a fixed j.
$\Omega_{j,i}$ for each $i \in [d]$
$Q_1, \ldots, Q_d \in \mathbb{R}^{n \times r}$ computed factor matrices
output: $\mathcal{H}$
for $i \in [d]$ **do**

    `# unfold measurements, mode-`i` fibers are columns,`
       `size `$m \times r^{(i-1)} m^{d-1-(i-1)}$
    $H \leftarrow \text{unfold}(\mathcal{B}_j, m \times r^{(i-1)} m^{d-1-(i-1)}, \text{mode} = i)$
    `# Undo the mode-`i` measurement operator and factor's`
       `action by finding least square solution to `$m \times r$
       `over-determined linear system`
    Solve $\Omega_{(j,i)} Q_i H_{\text{new}} = H$ for H_{new}
    `# reshape the flattened partially solved core into`
       `a tensor`
    $\mathcal{B}_j \leftarrow \text{fold}(H_{\text{new}}, \underbrace{r \times r \cdots \times r}_{i} \times \underbrace{m_c \times \cdots \times m_c}_{d-i}, \text{mode} = i)$
    `# Each iteration `$m \rightarrow r$` in `i`th mode or `$n \rightarrow n$` when`
       $i = j$

end

Algorithm 9 Recover HOSVD Kronecker One Pass.

```
# Obtain measurements for factors with Alg. 2
```
$B_1, B_2, \ldots, B_d \leftarrow$ Leave-One-Out Kronecker Sketching$\left(\left\{\Omega_{(i,j)}\right\}_{i,j\in[d]}, \mathcal{X}\right)$
```
# Obtain measurement for core with Alg. 4
```
$\mathcal{B}_c \leftarrow$ Core Sketching$(\{\Phi_i\}_{i\in[d]}, \mathcal{X})$
```
# Estimate factor matrices using Alg. 5
```
$Q_1, Q_2, \ldots, Q_d \leftarrow$ Recover HOSVD Factors from
Leave-One-Out Measurements $\left(\left\{\Omega_{(i,i)}\right\}_{i\in[d]}, \{B_i\}_{i\in[d]}, (r, \ldots, r)\right)$
```
# Estimate core using Alg. 8
```
$\mathcal{H} \leftarrow$ Recover HOSVD Core from Measurements$(\{\Phi_i\}_{i\in[d]}, \mathcal{B}_c)$
output: $\hat{\mathcal{X}} = [\![\mathcal{H}, Q_1, \ldots Q_d]\!]$

Algorithm 10 Recover HOSVD Kronecker Two Pass.

```
# Obtain measurements for factors with Alg. 2
```
$B_1, B_2, \ldots, B_d \leftarrow$ Leave-One-Out Kronecker Sketching$\left(\left\{\Omega_{(i,j)}\right\}_{i,j\in[d]}, \mathcal{X}\right)$
```
# Estimate factor matrices using Alg. 5
```
$Q_1, Q_2, \ldots, Q_d \leftarrow$ Recover HOSVD Factors from
Leave-One-Out Measurements$\left(\left\{\Omega_{(i,i)}\right\}_{i\in[d]}, \{B_i\}_{i\in[d]}, (r, \ldots, r)\right)$
```
# Compute core using Alg. 6
```
$\mathcal{H} \leftarrow$ Compute HOSVD Core with Second Access$(\{Q_i\}_{i\in[d]}, \mathcal{X})$
output: $\hat{\mathcal{X}} = [\![\mathcal{H}, Q_1, \ldots Q_d]\!]$

Algorithm 11 Recover HOSVD Khatri-Rao One Pass.

```
# Obtain measurements for factors with Alg. 3
```
$B_1, B_2, \ldots, B_d \leftarrow$ Leave-One-Out Khatri-Rao Sketching$\left(\left\{\Omega_{(i,j)}\right\}_{i,j\in[d]}, \mathcal{X}\right)$
```
# Obtain measurements for core with Alg. 4
```
$\mathcal{B}_c \leftarrow$ Core Sketching$(\{\Phi_i\}_{i\in[d]}, \mathcal{X})$
```
# Estimate factor matrices using Alg. 5
```
$Q_1, Q_2, \ldots, Q_d \leftarrow$ Recover HOSVD Factors from Leave-One-Out
Measurements $\left(\left\{\Omega_{(i,i)}\right\}_{i\in[d]}, \{B_i\}_{i\in[d]}, (r, \ldots, r)\right)$
```
# Estimate core using Alg. 8
```
$\mathcal{H} \leftarrow$ Recover HOSVD Core from Measurements$(\{\Phi_i\}_{i\in[d]}, \mathcal{B}_c)$
output: $\hat{\mathcal{X}} = [\![\mathcal{H}, Q_1, \ldots Q_d]\!]$

Algorithm 12 Recover HOSVD Khatri-Rao Two Pass.

```
# Obtain measurements for factors with Alg. 3
```
$B_1, B_2, \ldots, B_d \leftarrow$ Leave-One-Out Khatri-Rao Sketching($\left\{\Omega_{(i,j)}\right\}_{i,j\in[d]}, \mathcal{X}$)
```
# Estimate factor matrices using Alg. 5
```
$Q_1, Q_2, \ldots, Q_d \leftarrow$ Recover HOSVD Factors from Leave-One-Out
Measurements ($\left\{\Omega_{(i,i)}\right\}_{i\in[d]}, \{B_i\}_{i\in[d]}, (r, \ldots, r)$)
```
# Compute core using Alg. 6
```
$\mathcal{H} \leftarrow$ Compute HOSVD Core with Second Access($\{Q_i\}_{i\in[d]}, \mathcal{X}$)
output: $\hat{\mathcal{X}} = [\![\mathcal{H}, Q_1, \ldots Q_d]\!]$

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
2. Cichocki, A.: Tensor decompositions: new concepts in brain data analysis? J. Soc. Inst. Control Eng. **50**(7), 507–516 (2011)
3. Sun, J., Tao, D., Papadimitriou, S., Yu, P.S., Faloutsos, C.: Incremental tensor analysis: Theory and applications. ACM Trans. Knowl. Discov. Data. (TKDD) **2**(3), 1–37 (2008)
4. Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H.: Diffusion tensor imaging: concepts and applications. J. Magn. Reson. Imaging: An Official J. Int. Soc. Magn. Reson. Med. **13**(4), 534–546 (2001)
5. Chen, H., Ahmad, F., Vorobyov, S., Porikli, F.: Tensor decompositions in wireless communications and mimo radar. IEEE J. Sel. Top. Signal Process. **15**(3), 438–453 (2021)
6. Kaya, O., Uçar, B.: High performance parallel algorithms for the Tucker decomposition of sparse tensors. In: 2016 45th International Conference on Parallel Processing (ICPP), pp. 103–112 (2016). IEEE
7. Austin, W., Ballard, G., Kolda, T.G.: Parallel tensor compression for large-scale scientific data. In: 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 912–922 (2016). IEEE
8. Zhou, G., Cichocki, A., Xie, S.: Decomposition of big tensors with low multilinear rank. arXiv:1412.1885 (2014)
9. Tsourakakis, C.E.: Mach: Fast randomized tensor decompositions. In: Proceedings of the 2010 SIAM International Conference on Data Mining, pp. 689–700 (2010). SIAM
10. Malik, O.A., Becker, S.: Low-rank Tucker decomposition of large tensors using TensorSketch. Adv. Neural. Inf. Process. Syst. **31** (2018)
11. Rauhut, H., Schneider, R., Stojanac, Ž: Low rank tensor recovery via iterative hard thresholding. Linear Algebra Appl. **523**, 220–262 (2017)
12. Grotheer, R., Li, S., Ma, A., Needell, D., Qin, J.: Stochastic iterative hard thresholding for low-Tucker-rank tensor recovery. In: 2020 Information Theory and Applications Workshop (ITA), pp. 1–5 (2020). IEEE

13. Haselby, C.A., Iwen, M.A., Needell, D., Perlmutter, M., Rebrova, E.: Modewise operators, the tensor restricted isometry property, and low-rank tensor recovery. Appl. Comput. Harmon. Anal. **66**, 161–192 (2023)

14. Jiang, Q.: Near-isometric properties of Kronecker-structured random tensor embeddings. Adv. Neural. Inf. Process. Syst. **35**, 10191–10202 (2022)

15. Hendrikx, S., De Lathauwer, L.: Block row Kronecker-structured linear systems with a low-rank tensor solution. Front. Appl. Math. Stat. **8** (2022). https://doi.org/10.3389/fams.2022.832883

16. Sun, Y., Guo, Y., Luo, C., Tropp, J., Udell, M.: Low-rank Tucker approximation of a tensor from streaming data. SIAM J. Math. Data Sci. **2**(4), 1123–1150 (2020)

17. Yu, W., Gu, Y., Li, J., Liu, S., Li, Y.: Single-pass PCA of large high-dimensional data. arXiv:1704.07669 (2017)

18. Zare, A., Ozdemir, A., Iwen, M.A., Aviyente, S.: Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor pca. Proc. IEEE **106**(8), 1341–1358 (2018)

19. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)

20. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. Springer Series in Computational Mathematics: 56. Springer, Berlin (2019)

21. Muthukrishnan, S., et al.: Data streams: Algorithms and applications. Found. Trends® Theor. Comput. Sci. **1**(2), 117–236 (2005)

22. Minster, R., Saibaba, A.K., Kilmer, M.E.: Randomized algorithms for low-rank tensor decompositions in the Tucker format. SIAM J. Math. Data Sci. **2**(1), 189–215 (2020)

23. Johnson, W.B.: Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**, 189–206 (1984)

24. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Springer, New York (2013)

25. Baraniuk, R.G., Davenport, M.A., DeVore, R.A., Wakin, M.B.: A simple proof of the restricted isometry property for random matrices. Constructive Approximation (2007)

26. Sarlos, T.: Improved approximation algorithms for large matrices via random projections. In: 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), pp. 143–152 (2006). IEEE

27. Iwen, M.A., Needell, D., Rebrova, E., Zare, A.: Lower memory oblivious (tensor) subspace embeddings with fewer random bits: modewise methods for least squares. SIAM J. Matrix Anal. Appl. **42**(1), 376–416 (2021)

28. Magen, A., Zouzias, A.: Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In: Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1422–1436 (2011). SIAM

29. Cohen, M.B., Elder, S., Musco, C., Musco, C., Persu, M.: Dimensionality reduction for k-means clustering and low rank approximation. In: Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing, pp. 163–172 (2015)

30. Feldman, D., Schmidt, M., Sohler, C.: Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. SIAM J. Comput. **49**(3), 601–657 (2020)

31. Chowdhury, A., Yang, J., Drineas, P.: Structural conditions for projection-cost preservation via randomized matrix multiplication. Linear Algebra Appl. **573**, 144–165 (2019)

32. Musco, C., Musco, C.: Projection-cost-preserving sketches: Proof strategies and constructions. CoRR **abs/2004.08434** (2020)

33. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. SIAM J. Math. Anal. **43**(3), 1269–1281 (2011)

34. Jin, R., Kolda, T.G., Ward, R.: Faster Johnson-Lindenstrauss transforms via Kronecker products. Inf. Inference: J. IMA **10**(4), 1533–1562 (2021)

35. Bamberger, S., Krahmer, F., Ward, R.: Johnson-Lindenstrauss embeddings with Kronecker structure. SIAM J. Matrix Anal. Appl. **43**(4), 1806–1850 (2022)

36. Høgsgaard, M.M., Kamma, L., Larsen, K.G., Nelson, J., Schwiegelshohn, C.: Sparse dimensionality reduction revisited. arXiv (2023)

37. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. SIAM J. Matrix Anal. Appl. **21**(4), 1324–1342 (2000)

38. Vannieuwenhoven, N., Vandebril, R., Meerbergen, K.: A new truncation strategy for the higher-order singular value decomposition. SIAM J. Sci. Comput. **34**(2), 1027–1052 (2012)

39. Ahle, T.D., Kapralov, M., Knudsen, J.B., Pagh, R., Velingker, A., Woodruff, D.P., Zandieh, A.: Oblivious sketching of high-degree polynomial kernels. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 141–160 (2020). SIAM

40. Ahle, T.D., Knudsen, J.B.T.: Almost optimal tensor sketch. arXiv (2019)

41. Latała, R.: Estimation of moments of sums of independent real random variables. Ann. Probab. **25**(3), 1502–1513 (1997)

42. Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2018)

43. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes. Applied and numerical harmonic analysis. Springer, New York (2002)

44. Knyazev, A.V., Argentati, M.E.: Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. SIAM J. Sci. Comput. **23**(6), 2008–2040 (2002)

45. Woolfe, F., Liberty, E., Rokhlin, V., Tygert, M.: A fast randomized algorithm for the approximation of matrices. Appl. Comput. Harmon. Anal. **25**(3), 335–366 (2008)

46. Arriaga, R.I., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projection. Mach. Learn. **63**, 161–182 (2006)

Authors and Affiliations

Cullen Haselby[1] · Mark A. Iwen[1,2] · Deanna Needell[3] · Elizaveta Rebrova[4] · William Swartworth[3]

✉ Mark A. Iwen
iwenmark@msu.edu

Cullen Haselby
haselbyc@msu.edu

Deanna Needell
deanna@math.ucla.edu

Elizaveta Rebrova
elre@princeton.edu

William Swartworth
wswartworth@math.ucla.edu

[1] Department of Mathematics, Michigan State University, East Lansing, MI, USA

[2] Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

[3] Department of Mathematics, University of California, Los Angeles, CA, USA

[4] Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ, USA