Tensorized Random Projections

Beheshteh T. Rakhshan
Department of Mathematics, Purdue University

Guillaume Rabusseau*
DIRO and Mila, Université de Montréal

Abstract

We introduce a novel random projection technique for efficiently reducing the dimension of very high-dimensional tensors. Building upon classical results on Gaussian random projections and Johnson-Lindenstrauss transforms (JLT), we propose two tensorized random projection maps relying on the tensor train (TT) and CP decomposition format, respectively. The two maps offer very low memory requirements and can be applied efficiently when the inputs are low rank tensors given in the CP or TT format. Our theoretical analysis shows that the dense Gaussian matrix in JLT can be replaced by a low-rank tensor implicitly represented in compressed form with random factors, while still approximately preserving the Euclidean distance of the projected inputs. In addition, our results reveal that the TT format is substantially superior to CP in terms of the size of the random projection needed to achieve the same distortion ratio. Experiments on synthetic data validate our theoretical analysis and demonstrate the superiority of the TT decomposition.

1 Introduction

Random projections (RP) are commonly used in data science and machine learning to project down high-dimensional data into a lower dimensional space while preserving most of the relevant information in the data [Vempala, 2005, Bingham and Mannila, 2001]. These methods have been successfully used to trade accuracy in order to reduce time and storage complexity of classical learning algorithms such as k-nearest neighbors [Ailon and Chazelle, 2006, 2009, Indyk and Motwani, 1998, Kleinberg, 1997], k-means [Boutsidis et al., 2010], support vector machines [Paul et al., 2013] and learning high-dimensional Gaussian mixtures [Dasgupta, 1999, 2000] to name a few. Most modern RP techniques build upon the celebrated *Johnson-Lindenstrauss lemma* [Johnson and Lindenstrauss, 1984] which shows that an arbitrary number of high-dimensional points can be linearly projected into an exponentially lower dimensional subspace while preserving distances between points. One of the simplest Johnson-Lindenstrauss transforms (JLT) is constructed from a random matrix $\mathbf{A}$ whose entries are independently and identically drawn from a normal distribution. Fast variants of JLT have been proposed by introducing sparsity in $\mathbf{A}$ [Achlioptas, 2003, Li et al., 2006] and by leveraging fast matrix multiplication algorithms [Ailon and Chazelle, 2006, 2009, Ailon and Liberty, 2013].

At the same time, tensor decomposition techniques have also recently emerged as a powerful tool for dealing with high-dimensional data. Tensor methods are particularly suited to handle high-dimensional multi-modal data and have been successfully applied in neuroimaging [Zhou et al., 2013], signal processing [Cichocki et al., 2009, Sidiropoulos et al., 2017], spatio-temporal analysis [Bahadori et al., 2014] and computer vision [Lu et al., 2013]. But even when the data is not inherently multi-modal in nature, tensor decomposition techniques can be used to speed-up and scale classical learning algorithms to very high-dimensional spaces [Novikov et al., 2014, 2015]. Such algorithms exploit the ability of tensor decomposition techniques to implicitly represent very high-dimensional data in compressed form, by first tensorizing the data before applying tensor decomposition techniques. In particular, the CANDECOMP/PARAFAC (CP) [Hitchcock, 1927] and tensor-train (TT) [Oseledets, 2011] decomposition can represent Nth order d-dimensional tensors (or equivalently d^N-dimensional vectors) using only $\mathcal{O}(NdR)$ and $\mathcal{O}(NdR^2)$ parameters respectively, where the rank parameter R controls the coarseness of the decomposition. Crucially, the number of parameters of these

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

* CIFAR AI Chair

decomposition only grows *linearly* with the order of the tensor N, which is not the case for other popular decomposition models such as the Tucker decomposition [Tucker, 1966].

While efficient random projection techniques have been proposed to deal with high-dimensional data, RP still suffer from the curse of dimensionality when the input dimension is very large, which is the case for high-order tensor inputs. In this work, we propose to leverage tensor decomposition techniques to *tensorize* Gaussian random projections. In doing so, we design efficient random projections that can be applied to any high-order tensor inputs with arbitrary rank and structure. In particular, projecting an input tensor given in the CP or TT format can be done very efficiently. More precisely, we propose two tensorized random projection maps, $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$, relying on the TT and CP formats respectively.

Intuitively, the random projection maps $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$ are constructed by enforcing a low rank tensor structure (CP or TT) on the rows of the random projection matrix $\mathbf{A} \in \mathbb{R}^{k \times d^N}$ where $k \ll d^N$ is the size of the random projection and the inputs are Nth-order d-dimensional tensors. The parameter R corresponds to the rank of the CP/TT decomposition used to represent the rows of $\mathbf{A}$ and controls the tradeoff between the quality of the embedding and the computational and memory cost of projecting input points. More precisely, if the input $\boldsymbol{\mathcal{X}}$ is given as a rank $\tilde{R}$ CP or TT tensor, computing $f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})$ and $f_{\mathrm{CP}(R)}(\boldsymbol{\mathcal{X}})$ can be done in time $\mathcal{O}(kNd\max(R,\tilde{R})^3)$. In terms of memory requirements, $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$ have $\mathcal{O}(kNdR^2)$ and $\mathcal{O}(kNdR)$ parameters respectively. In comparison the cost of transformation for a Gaussian JLT is in $\mathcal{O}(kd^N)$ which can be improved to $\mathcal{O}(k + Nd^N \log d)$ using fast JLT.

Our theoretical analysis shows that the key properties of Gaussian random projections are preserved after *tensorization*: for any $\varepsilon > 0$, with high probability, our tensorized RP embed any set of m points up to multiplicative distortion $(1 \pm \varepsilon)$ as soon as $k \gtrsim \varepsilon^{-2}(1+2/R)^N \log^{2N} m$ for $f_{\mathrm{TT}(R)}$ and $k \gtrsim \varepsilon^{-2} 3^{N-1}(1 + 2/R)\log^{2N} m$ for $f_{\mathrm{CP}(R)}$. Besides showing that both tensorizations lead to efficient random projections (in terms of time and memory complexity), our analysis further reveals that $f_{\mathrm{TT}(R)}$ is substantially superior to $f_{\mathrm{CP}(R)}$ in terms of the size of the random projection needed to achieve the same multiplicative distortion. This can be seen by comparing the exponential dependency on the order N of input tensors in the lower bounds on k given above (and how increasing the rank R of the tensorized map can mitigate this dependency). In particular, our analysis shows that the CP format is not a reasonable decomposition format for tensorizing random projections in the case of high order input tensors.

Summary of contributions. We present two tensorized random projection maps, $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$, using the TT and CP decomposition models respectively. We show that both maps are *Johnson-Lindenstrauss Transforms* offering appealing computational and memory requirements. In particular, our work is the first to design efficient RP for input tensors given in the CP or TT format. Our theoretical analysis for $f_{\mathrm{CP}(R)}$ extends the one first initiated in [Sun et al., 2018] (which was limited to matrix inputs) to high-order input tensors. To the best of our knowledge, this is the first time that the TT decomposition model is leveraged to design RP that can scale to very high-dimensional inputs. Our theoretical analysis further shows that the TT format is a better decomposition model than CP for tensorizing random projection maps. Our numerical simulations substantially validate this conclusion. It is worth mentioning that our analysis is not focused on rank-one tensors and holds for arbitrary input tensors with low CP rank or TT rank structure.

Related work. Tensor Sketch [Pham and Pagh, 2013] is an extension of the Count Sketch algorithm [Charikar et al., 2002] using fast FFT which can efficiently approximate polynomial kernels. More recently, [Shi and Anandkumar, 2019] extended Tensor Sketch to exploit the multi-modal structure of tensor inputs, but their approach relies on the Tucker decomposition format and cannot scale to very high-order tensors. Kapralov et al. [Ahle et al., 2020] also consider sketching tensor products of data points without explicitly forming the resulting tensor, and propose an algorithm to compute a linear sketch for degree-N polynomial kernels.

More closely related to our work, Sun et al. [Sun et al., 2018] introduce a Tensor Random Projection map (TRP) using a row-wise Kronecker product of random matrices. We show that their method is equivalent to the CP tensorized random projection map studied in this paper. Their theoretical analysis is limited to order 2 tensors (i.e. matrices) and rank one projection maps: they show that TRP satisfies the JL property when $k \gtrsim \varepsilon^{-2}\log^8 m$ for $N = 2$ and $R = 1$. Our results for $f_{\mathrm{CP}(R)}$ extend theirs to arbitrary values of N and R and provide tighter bounds even for the case of $N = 2$ and $R = 1$.

Lastly, Jin et al. [Jin et al., 2019] extend the fast JLT for embedding vectors with a Kronecker product structure. They show that the map they propose satisfy the JL property when $k \gtrsim \varepsilon^{-2}\log^{2N-1} m \log(d^N)$ (up to polylog factors) and that projecting a rank one tensor can be done in $\mathcal{O}(Nd\log d + k)$. While the upper bound we derive for $f_{\mathrm{TT}(R)}$ is comparable, the choice of the rank parameter gives more flexibility to control the trade-off between accuracy and computational efficiency. In particular, com-

puting $f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})$ can be considerably faster than the method proposed in [Jin et al., 2019] when $\boldsymbol{\mathcal{X}}$ is a low rank tensor given in the TT format (see Section 4.1).

2 Preliminaries

In this section, we introduce our notations and present the necessary background on tensor algebra, tensor decomposition and random projections. More details can be found in [Kolda and Bader, 2009, Vempala, 2005, Dasgupta and Gupta, 2003].

2.1 Notations

We use lower case bold letters for vectors (*e.g.* $\mathbf{a}, \mathbf{b}$, ...), upper case bold letters for matrices (*e.g.* $\mathbf{A}, \mathbf{B}$, ...), and bold calligraphic letters for higher order tensors (*e.g.* $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$, ...). If $\mathbf{v} \in \mathbb{R}^{d_1}$ and $\mathbf{u} \in \mathbb{R}^{d_2}$, we use $\mathbf{v} \otimes \mathbf{u} \in \mathbb{R}^{d_1 d_2}$ to denote the Kronecker product between vectors. The 2-norm of a vector $\mathbf{u}$ is denoted by $\|\mathbf{u}\|_2$ or simply $\|\mathbf{u}\|$. The Khatri-Rao product is defined as the "matching column-wise" Kronecker product: if $\mathbf{A} \in \mathbb{R}^{m \times R}$ and $\mathbf{B} \in \mathbb{R}^{n \times R}$, it is denoted by $\mathbf{A} \odot \mathbf{B}$ and given by $[\mathbf{a}_1 \otimes \mathbf{b}_1 \cdots \mathbf{a}_R \otimes \mathbf{b}_R] \in \mathbb{R}^{mn \times R}$. We use the symbol "$\circ$" to denote the outer product (or tensor product) between vectors. Given a matrix $\mathbf{S} \in \mathbb{R}^{d_1 \times d_2}$, we use $\mathrm{vec}(\mathbf{S}) \in \mathbb{R}^{d_1 . d_2}$ to denote the column vector obtained by concatenating the columns of $\mathbf{S}$. The $d \times d$ identity matrix will be written as $\mathbf{I}_d$ and the transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^\mathsf{T}$. For any integer k we use $[k]$ to denote the set of integers from 1 to k. For scalars $x, y \in \mathbb{R}$, we use $x \gtrsim y$ to denote that $x \geq cy$ for some constant c.

2.2 Tensors

A *N-th order tensor* $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ can simply be seen as a multidimensional array ($\boldsymbol{\mathcal{S}}_{i_1, \cdots, i_N} : i_n \in [d_n], n \in [N]$). The inner product between tensors is defined by $\langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}} \rangle = \sum_{i_1, \cdots, i_N} \boldsymbol{\mathcal{S}}_{i_1, \cdots, i_N} \boldsymbol{\mathcal{T}}_{i_1, \cdots, i_N}$ for $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ and the Frobenius norm is defined by $\|\boldsymbol{\mathcal{S}}\|_F^2 = \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{S}} \rangle$. If $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{J_1 \times \cdots \times J_N}$, we use $\boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{B}} \in \mathbb{R}^{I_1 J_1 \times \cdots \times I_N J_N}$ to denote the Kronecker product of tensors. The *mode-n* fibers of $\boldsymbol{\mathcal{S}}$ are the vectors obtained by fixing all indices except the nth one. The *n-th mode matricization* of $\boldsymbol{\mathcal{S}}$ is the matrix having the mode-n fibers of $\boldsymbol{\mathcal{S}}$ for columns* and is denoted by $\boldsymbol{\mathcal{S}}_{(n)} \in \mathbb{R}^{d_n \times d_1 \cdots d_{n-1} d_{n+1} \cdots d_N}$. The vectorization of a tensor is the vector obtained by concatenating its mode-1 fibers, i.e., $\mathrm{vec}(\boldsymbol{\mathcal{S}}) = \mathrm{vec}(\boldsymbol{\mathcal{S}}_{(1)})$. The notion of matricization can be extended to any subset $I \subset [N]$ of the modes of $\boldsymbol{\mathcal{S}}$, resulting in a matrix $\boldsymbol{\mathcal{S}}_{(I)}$ of size $\prod_{i \in I} d_i \times \prod_{j \in [N] \setminus I} d_j$.

A rank R *CP decomposition* of a tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ consists in factorizing $\boldsymbol{\mathcal{S}}$ into a sum of R rank one tensors: $\boldsymbol{\mathcal{S}} = \sum_{r=1}^{R} \mathbf{a}_r^1 \circ \mathbf{a}_r^2 \circ \cdots \mathbf{a}_r^N$ where each $\mathbf{a}_r^n \in \mathbb{R}^{d_n}$. Stacking the vectors $\mathbf{a}_1^n, \ldots, \mathbf{a}_R^n$ into a factor matrix $\mathbf{A}^n \in \mathbb{R}^{d_n \times R}$ for each $n \in [N]$, we will concisely denote the CP decomposition by $\boldsymbol{\mathcal{S}} = [\![\mathbf{A}^1, \cdots, \mathbf{A}^N]\!]$.

A rank R *tensor train decomposition* of a tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ consists in factorizing $\boldsymbol{\mathcal{S}}$ into the the product of N 3rd-order core tensors $\boldsymbol{\mathcal{G}}^1 \in \mathbb{R}^{1 \times d_1 \times R}, \boldsymbol{\mathcal{G}}^2 \in \mathbb{R}^{R \times d_2 \times R}, \cdots, \boldsymbol{\mathcal{G}}^{N-1} \in \mathbb{R}^{R \times d_{N-1} \times R}, \boldsymbol{\mathcal{G}}^N \in \mathbb{R}^{R \times d_N \times 1}$, and is defined† by $\boldsymbol{\mathcal{S}}_{i_1, \cdots, i_N} = (\boldsymbol{\mathcal{G}}^1)_{i_1,:} (\boldsymbol{\mathcal{G}}^2)_{:,i_2,:} \cdots (\boldsymbol{\mathcal{G}}^{N-1})_{:,i_{N-1},:} (\boldsymbol{\mathcal{G}}^N)_{:,i_N}$, for all indices $i_1 \in [d_1], \cdots, i_N \in [d_N]$; we will use the notation $\boldsymbol{\mathcal{S}} = \langle\!\langle \boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^{N-1}, \boldsymbol{\mathcal{G}}^N \rangle\!\rangle$ to denote the TT decomposition.

2.3 Johnson-Lindenstrauss Transform

A classical result of Johnson-Lindenstrauss (JL) [Johnson and Lindenstrauss, 1984] states that any m-point set P in d dimension can be linearly projected to $k = \Omega(\varepsilon^{-2} \log(m))$ dimensions while approximately preserving the pairwise distances between the points. More precisely, there exists a map $f : \mathbb{R}^d \to \mathbb{R}^k (d \gg k)$ such that for all $\mathbf{u}, \mathbf{v} \in P$,

$$(1-\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1+\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2.$$

We will call a map satisfying this property a *Johnson-Lindenstrauss transform* (JLT). One of the simplest examples of a JL transform is the so-called *Gaussian random projection* map $f : \mathbf{x} \mapsto \frac{1}{\sqrt{k}} \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{k \times d}$ is a random matrix whose entries are independently drawn from a normal distribution. For a fixed set of input points in $\mathbb{R}^d$, f will satisfy the JL property with high probability. To cope with the computational cost and storage requirements of Gaussian random projections, sparse and very-sparse random projections were proposed in [Achlioptas, 2003] and [Li et al., 2006] respectively. These maps leverage the fact that the JL property is preserved even if only a small subset of the entries of $\mathbf{A}$ are normal variables while the other ones are set to 0.

It is easy to see that in order to be a JL transform, a map f must satisfy two fundamental properties: (i) it has to be an expected isometry, i.e. $\mathbb{E}\left[\|f(\mathbf{x})\|^2\right] = \|\mathbf{x}\|^2$, and (ii) the variance of $\|f(\mathbf{x})\|^2$ should quickly decrease to 0 as the size of the random projection k increases.

*The specific ordering of the fibers does not matter as long as it is consistent across all reshaping operations.

†The general definition of the TT-decomposition allows the rank R to be different for each mode, but this definition is sufficient for the purpose of this paper.

3 Tensorized Random Projections

As mentioned in the previous section, sparse and very-sparse Gaussian RP reduce time and memory complexity by enforcing the rows of the matrix $\mathbf{A}$ in the Gaussian RP $f : \mathbf{x} \to \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}$ to be sparse. In this work, we propose to enforce a low rank tensor structure on the rows of $\mathbf{A}$ instead to obtain better scalability w.r.t. the input dimension, which is crucial when dealing with high-order tensor inputs.

We present two tensorized random projection maps, $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$, relying on the TT and CP decomposition respectively. These maps embed any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ into $\mathbb{R}^k$, where $k \ll d_1 d_2 \cdots d_N$. Considering the case $d_1 = \cdots = d_N = d$ for simplicity, classical random projection maps would require $\mathcal{O}(kd^N)$ parameters (or $\mathcal{O}(k\sqrt{d^N})$ with very sparse random projections) which is costly when N is large. In contrast, $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$ only require $\mathcal{O}(kNdR^2)$ and $\mathcal{O}(kNdR)$ parameters respectively. The two maps are constructed similarly: each component of the projection is given by the inner product between the input and a random tensor with a low rank structure (w.r.t. either the TT or CP decomposition format). Formally, we have the following two definitions:

Definition 1. *A* TT random projection of rank R *is a linear map* $f_{\mathrm{TT}(R)} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^k$ *defined componentwise by*

$$\left(f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right)_i := \frac{1}{\sqrt{k}} \langle\langle \boldsymbol{\mathcal{G}}_i^1, \boldsymbol{\mathcal{G}}_i^2, \cdots, \boldsymbol{\mathcal{G}}_i^N \rangle\rangle, \boldsymbol{\mathcal{X}} \rangle, \ i \in [k]$$

where $\boldsymbol{\mathcal{G}}_i^1 \in \mathbb{R}^{1 \times d_1 \times R}, \boldsymbol{\mathcal{G}}_i^2 \in \mathbb{R}^{R \times d_2 \times R}, \cdots, \boldsymbol{\mathcal{G}}_i^{N-1} \in \mathbb{R}^{R \times d_{N-1} \times R}, \boldsymbol{\mathcal{G}}_i^N \in \mathbb{R}^{R \times d_N \times 1}$ *for* $i \in [k]$*, and the entries of each* $\boldsymbol{\mathcal{G}}_i^n$ *for* $i \in [k]$*,* $n \in [N]$ *are drawn independently from a Gaussian distribution with mean* 0 *and variance* $\frac{1}{\sqrt{R}}$ *if* $n \in \{1, N\}$ *and variance* $\frac{1}{R}$ *if* $1 < n < N$*.*

Definition 2. *A* CP random projection of rank R *is a linear map* $f_{\mathrm{CP}(R)} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^k$ *defined componentwise by*

$$\left(f_{\mathrm{CP}(R)}(\boldsymbol{\mathcal{X}})\right)_i := \frac{1}{\sqrt{k}} \langle [\![\mathbf{A}_i^1, \mathbf{A}_i^2, \cdots, \mathbf{A}_i^N]\!], \boldsymbol{\mathcal{X}} \rangle, \ i \in [k]$$

where each $\mathbf{A}_i^n \in \mathbb{R}^{d_n \times R}$ *for* $i \in [k]$*,* $n \in [N]$ *and the entries of each* $\mathbf{A}_i^n$ *are drawn independently from a Gaussian distribution with mean* 0 *and variance* $\left(\frac{1}{R}\right)^{\frac{1}{N}}$*.*

One can check that applying these projection maps on an input tensor given in the CP or the TT format can be done efficiently: the complexity of computing $f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})$ is in $\mathcal{O}(kNd \max(R, \tilde{R})^3)$ if $\boldsymbol{\mathcal{X}}$ is given as a rank $\tilde{R}$ CP or TT tensor, and the complexity for $f_{\mathrm{CP}(R)}(\boldsymbol{\mathcal{X}})$ is in $\mathcal{O}(kNd \max(R, \tilde{R})^2)$ if $\boldsymbol{\mathcal{X}}$ is in the CP format and in $\mathcal{O}(kNd \max(R, \tilde{R})^3)$ if $\boldsymbol{\mathcal{X}}$ is in the TT format (where we assumed $d_1 = \cdots = d_N = d$ for simplicity).

Before studying the properties of these tensorized random projections in the next section, we show how $f_{\mathrm{CP}(\cdot)}$ is equivalent to the tensor random projection map proposed in [Sun et al., 2018]. In this work, the authors introduce the map

$$f_{\mathrm{TRP}}(\boldsymbol{\mathcal{X}}) := \frac{1}{\sqrt{k}} (\mathbf{A}^1 \odot \mathbf{A}^2 \odot \cdots \odot \mathbf{A}^N)^\mathsf{T} \mathrm{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^k,$$

where each $\mathbf{A}^n \in \mathbb{R}^{d_n \times k}$ for $n \in [N]$ is a random matrix whose entries are i.i.d random variables with mean zero and variance one. One can check that f_{TRP} is strictly equivalent to $f_{\mathrm{CP}(1)}$ using basic properties of the CP decomposition. Furthermore, the authors introduce a variance reduction technique with the map $f_{\mathrm{TRP}(T)}$, a scaled average of T independent TRPs, defined by $f_{\mathrm{TRP}(T)}(\boldsymbol{\mathcal{X}}) := \frac{1}{\sqrt{T}} \sum_{t=1}^T f_{\mathrm{TRP}}^{(t)}(\boldsymbol{\mathcal{X}})$. Again, one can easily check the strict equivalence between $f_{\mathrm{CP}(R)}$ and $f_{\mathrm{TRP}(T)}$ when $R = T$.

4 Main Results

In this section, we present our main results showing that the tensorized projection $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$ still benefits from the fundamental properties of Gaussian random projections: they are expected isometry and the variance of the norm of the projections decreases to 0 as the embedding dimension k grows. These results imply that, in addition to be particularly efficient in terms of storage requirement and computational cost, these maps are JL transforms: they approximately preserve Euclidean distances between projected points. Moreover, our analysis will show that there is a crucial difference between the two tensorized random projections: as the order of the input tensor $\boldsymbol{\mathcal{X}}$ grows, the embedding dimension of $f_{\mathrm{CP}(R)}$ needs to grow exponentially in comparison to the one of $f_{\mathrm{TT}(R)}$ in order to achieve the same distortion ratio ε. Our results rely on the following theorem which shows that both maps are expected isometries and gives bounds on the variance of the two projections.

Theorem 1. *Let* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$*. The random projection maps* $f_{\mathrm{TT}(R)}$ *and* $f_{\mathrm{CP}(R)}$ *(see Definitions 1 and 2) satisfy the following properties:*

- $\mathbb{E}\left[\|f_{\mathrm{CP}(R)}(\boldsymbol{\mathcal{X}})\|_2^2\right] = \mathbb{E}\left[\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\|_2^2\right] = \|\boldsymbol{\mathcal{X}}\|_F^2$
- $\mathrm{Var}\left(\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\|_2^2\right) \leq \frac{1}{k}\left(3\left(1 + \frac{2}{R}\right)^{N-1} - 1\right)\|\boldsymbol{\mathcal{X}}\|_F^4$
- $\mathrm{Var}\left(\|f_{\mathrm{CP}(R)}(\boldsymbol{\mathcal{X}})\|_2^2\right) \leq \frac{1}{k}\left(3^{N-1}\left(1 + \frac{2}{R}\right) - 1\right)\|\boldsymbol{\mathcal{X}}\|_F^4$

The proof of this theorem for the TT random projection map is given in the next section and the proof for $f_{\mathrm{CP}(R)}$ can be found in the Appendix.

In the case of vector inputs, *i.e.* $N = 1$, we recover the classical expression for the variance of Gaussian random projections given by $\text{Var}\left(\|f(\mathbf{x})\|\right)^2 = \frac{2}{k}\|\mathbf{x}\|^4$ (note that in this setting R is necessarily equal to 1 since $N = 1$).

It is worth mentioning that the only inequality used to derive the bounds comes from the sub-multiplicativity of the Frobenius norm applied to matricizations of the input tensor $\boldsymbol{\mathcal{X}}$. For example, for the case of order 2 input tensors, *i.e.* matrices, the variance of $f_{\text{TT}(R)}$ is given by

$$\text{Var}\left(\|f_{\text{TT}(R)}(\mathbf{X})\|^2\right) = \frac{1}{k}\left(2\|\mathbf{X}\|_F^4 + \frac{6}{R}\text{Tr}\left[(\mathbf{X}^\top\mathbf{X})^2\right]\right).$$

Comparing now the bounds on the variance of $f_{\text{TT}(R)}$ and $f_{\text{CP}(R)}$, we observe that while both bounds have an exponential dependency on the order N of the input tensors, slightly increasing the rank R of the TT random projection mitigates this dependency while it has no effect for the CP random projection. This shows that $f_{\text{CP}(R)}$ is not a suitable RP since k has to grow exponentially in N in order to approach the variance of classical Gaussian random projections. Using the bounds on the variance of the projections, we can now derive lower bounds on the size k of the random projections $f_{\text{CP}(R)}$ and $f_{\text{TT}(R)}$ needed to satisfy the JL property with high probability.

Theorem 2. *Let $P \subset \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ be a set of m order N tensors. Then, for any $\varepsilon > 0$ and any $\delta > 0$, the following hold simultaneously for all $\boldsymbol{\mathcal{X}} \in P$:*

- *if $k \gtrsim \varepsilon^{-2}(1+2/R)^N \log^{2N}\left(\frac{m}{\delta}\right)$ then $\mathbb{P}(\left\|f_{\text{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 = (1\pm\varepsilon)\|\boldsymbol{\mathcal{X}}\|_F^2) \geq 1-\delta$,*
- *if $k \gtrsim \varepsilon^{-2}3^{N-1}(1+2/R)\log^{2N}\left(\frac{m}{\delta}\right)$ then $\mathbb{P}(\left\|f_{\text{CP}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 = (1\pm\varepsilon)\|\boldsymbol{\mathcal{X}}\|_F^2) \geq 1-\delta$.*

4.1 Comparison to related work

We conclude this section by comparing the previous theorem with the closest related work. Jin et al. [Jin et al., 2019] proposed a Kronecker structured JL transform satisfying the JL property for m points with probability $1-\delta$ as soon as $k \gtrsim \varepsilon^{-2}\log^{2N-1}\left(\frac{m}{\delta}\right)\log(d^N)$, up to polylog factors. Our results are similar to theirs but differ in one key aspect. In their work, projecting a rank one tensor can be done in $\mathcal{O}(Nd\log d + k)$. Hence, by linearity, projecting a tensor of rank $\tilde{R}$ given in the CP format can be done in $\mathcal{O}(\tilde{R}(Nd\log d + k))$. However, low rank tensors given in the TT format cannot be efficiently projected using their method[‡], In contrast, $f_{\text{TT}(R)}(\boldsymbol{\mathcal{X}})$ and $f_{\text{CP}(R)}(\boldsymbol{\mathcal{X}})$ can both be computed in $\mathcal{O}(kNd\max(R,\tilde{R})^3)$ when $\boldsymbol{\mathcal{X}}$ is given as a rank $\tilde{R}$ CP or TT tensor; our approach is thus better suited for inputs given in the TT format. In [Sun et al., 2018], they proposed a tensor random projection map for sub-Gaussian random variables. They give a lower bound of $k \gtrsim \varepsilon^{-2}\log^8\left(\frac{m}{\delta}\right)$ only for the case of order 2 input tensors, treating the rank parameter R as a constant. Moreover, even in the case of order 2 input tensors our lower bound of $\varepsilon^{-2}(1+2/R)\log^4\left(\frac{m}{\delta}\right)$ is tighter than the one they provide.

[‡]Indeed, almost all low rank TT tensors have exponentially large CP rank (see e.g, Theorem 1 in [Khrulkov et al., 2018])

5 Proofs

In this section, we present the proofs of our results for the random projection map $f_{\text{TT}(R)}$. The techniques used for the map $f_{\text{CP}(R)}$ are of a similar flavor and can be found in the Appendix.

5.1 Proof of Theorem 1: TT case

Expected isometry. We start by showing that $f_{\text{TT}(R)}$ is an expected isometry, *i.e.* that $\mathbb{E}\left\|f_{\text{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 = \|\boldsymbol{\mathcal{X}}\|_F^2$. Let $y_i = \langle\langle\boldsymbol{\mathcal{G}}_i^1, \boldsymbol{\mathcal{G}}_i^2, \cdots, \boldsymbol{\mathcal{G}}_i^N\rangle\rangle, \boldsymbol{\mathcal{X}}\rangle$ and $\mathbf{y} = [y_1, y_2, \cdots, y_k]$. With these definitions we have $f_{\text{TT}(R)}(\boldsymbol{\mathcal{X}}) = \frac{1}{\sqrt{k}}\mathbf{y}$ and it is thus sufficient to find $\mathbb{E}[y_1^2]$. To lighten the notation, let $\boldsymbol{\mathcal{G}}^n = \boldsymbol{\mathcal{G}}_1^n$ for each $n \in [N]$ and let $\boldsymbol{\mathcal{S}} = \langle\langle\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^N\rangle\rangle$. We have

$$\begin{aligned}\mathbb{E}[y_1^2] &= \mathbb{E}[\langle\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}\rangle^2] = \mathbb{E}[\langle\boldsymbol{\mathcal{S}}\otimes\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}\otimes\boldsymbol{\mathcal{X}}\rangle] \\ &= \langle\mathbb{E}[\boldsymbol{\mathcal{S}}\otimes\boldsymbol{\mathcal{S}}], \boldsymbol{\mathcal{X}}\otimes\boldsymbol{\mathcal{X}}\rangle.\end{aligned}$$

Using the fact that the core tensors $\boldsymbol{\mathcal{G}}^n$ are independent, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mathcal{S}}\otimes\boldsymbol{\mathcal{S}}] &= \mathbb{E}[\langle\langle\boldsymbol{\mathcal{G}}^1\otimes\boldsymbol{\mathcal{G}}^1, \cdots, \boldsymbol{\mathcal{G}}^N\otimes\boldsymbol{\mathcal{G}}^N\rangle\rangle] \\ &= \langle\langle\mathbb{E}[\boldsymbol{\mathcal{G}}^1\otimes\boldsymbol{\mathcal{G}}^1], \cdots, \mathbb{E}[\boldsymbol{\mathcal{G}}^N\otimes\boldsymbol{\mathcal{G}}^N]\rangle\rangle.\end{aligned}$$

Now, for $1 < n < N$, since the entries of each core tensor $\boldsymbol{\mathcal{G}}^n$ are i.i.d. Gaussian variables with mean 0 and variance $1/R$, we have

$$\mathbb{E}[\boldsymbol{\mathcal{G}}^n\otimes\boldsymbol{\mathcal{G}}^n] = \frac{1}{R}\text{vec}(\mathbf{I}_R)\circ\text{vec}(\mathbf{I}_{d_n})\circ\text{vec}(\mathbf{I}_R).$$

Similarly, $\mathbb{E}[\boldsymbol{\mathcal{G}}^1\otimes\boldsymbol{\mathcal{G}}^1] = \frac{1}{\sqrt{R}}\text{vec}(\mathbf{I}_{d_1})\circ\text{vec}(\mathbf{I}_R)$ and $\mathbb{E}[\boldsymbol{\mathcal{G}}^N\otimes\boldsymbol{\mathcal{G}}^N] = \frac{1}{\sqrt{R}}\text{vec}(\mathbf{I}_R)\circ\text{vec}(\mathbf{I}_{d_N})$.

A careful but straightforward derivation consequently shows that $\mathbb{E}[\boldsymbol{\mathcal{S}}\otimes\boldsymbol{\mathcal{S}}] = \text{vec}(\mathbf{I}_{d_1})\circ\cdots\circ\text{vec}(\mathbf{I}_{d_N})$, which implies $\mathbb{E}[y_1^2] = \langle\mathbb{E}[\boldsymbol{\mathcal{S}}\otimes\boldsymbol{\mathcal{S}}], \boldsymbol{\mathcal{X}}\otimes\boldsymbol{\mathcal{X}}\rangle = \|\boldsymbol{\mathcal{X}}\|_F^2$. From which $\mathbb{E}\left\|f_{\text{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 = \|\boldsymbol{\mathcal{X}}\|_F^2$ directly follows.

Bound on the variance of $f_{\text{TT}(R)}$. In order to bound the variance of $\|\mathbf{y}\|_2^2$ we need to bound $\mathbb{E}[\|\mathbf{y}\|_2^4]$. We have

$$\mathbb{E}[\|\mathbf{y}\|_2^4] = \sum_{i=1}^k \mathbb{E}[y_i^4] + \sum_{i\neq j}\mathbb{E}[y_i^2 y_j^2].$$

Since y_i and y_j are independent whenever $i \neq j$ and $\mathbb{E}[y_i^2] = \|\boldsymbol{\mathcal{X}}\|_F^4$ for all i, the second summand is equal to $k(k-1)\|\boldsymbol{\mathcal{X}}\|_F^4$. We now derive a bound on $\mathbb{E}[y_1^4]$.

Our proof relies on the following technical lemmas. The first one is a direct consequence of *Isserlis' theorem* [Isserlis, 1918] and the second one follows from standard properties of the Wishart distribution (see *e.g.* Section 3.3.6 of [Gupta and Nagar, 2018]).

Lemma 3. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a random matrix whose entries are i.i.d normal random variables with mean zero and variance σ^2, and let $\mathbf{B} \in \mathbb{R}^{m\times n}$ be a (random) matrix independent of $\mathbf{A}$. Then,*

$$\mathbb{E}\langle \mathbf{A}, \mathbf{B}\rangle^4 = 3\sigma^4 \mathbb{E}\,\|\mathbf{B}\|_F^4 .$$

Proof. Setting $\mathbf{a} = \mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$ and $\mathbf{b} = \mathrm{vec}(\mathbf{B}) \in \mathbb{R}^{mn}$, we have

$$\begin{aligned}\mathbb{E}\langle \mathbf{A}, \mathbf{B}\rangle^4 &= \mathbb{E}\langle \mathbf{a}, \mathbf{b}\rangle^4 \\ &= \mathbb{E}\langle \mathbf{a}^{\otimes 4}, \mathbf{b}^{\otimes 4}\rangle = \langle \mathbb{E}[\mathbf{a}^{\otimes 4}], \mathbb{E}[\mathbf{b}^{\otimes 4}]\rangle,\end{aligned}$$

where the last equality is obtained by using the independence between $\mathbf{a}$ and $\mathbf{b}$. Element-wise, by using Isserlis' theorem [Isserlis, 1918] and using the fact that $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ we have,

$$\begin{aligned}(\mathbb{E}[\mathbf{a}^{\otimes 4}])_{i_1,i_2,i_3,i_4} &= \mathbb{E}[\mathbf{a}_{i_1}\mathbf{a}_{i_2}\mathbf{a}_{i_3}\mathbf{a}_{i_4}] \\ &= \mathbb{E}[\mathbf{a}_{i_1}\mathbf{a}_{i_2}]\mathbb{E}[\mathbf{a}_{i_3}\mathbf{a}_{i_4}] + \mathbb{E}[\mathbf{a}_{i_1}\mathbf{a}_{i_3}]\mathbb{E}[\mathbf{a}_{i_2}\mathbf{a}_{i_4}] \\ &\quad + \mathbb{E}[\mathbf{a}_{i_1}\mathbf{a}_{i_4}]\mathbb{E}[\mathbf{a}_{i_2}\mathbf{a}_{i_3}] \\ &= (\delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3})\sigma^4,\end{aligned}$$

where δ is the Kronecker symbol. Therefore, letting $\Delta_{i_1i_2i_3i_4} = \delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3}$, we obtain

$$\begin{aligned}\mathbb{E}\langle \mathbf{A}, \mathbf{B}\rangle^4 &= \sum_{i_1,i_2,i_3,i_4} \mathbb{E}[\mathbf{a}_{i_1}\mathbf{a}_{i_2}\mathbf{a}_{i_3}\mathbf{a}_{i_4}]\mathbb{E}[\mathbf{b}_{i_1}\mathbf{b}_{i_2}\mathbf{b}_{i_3}\mathbf{b}_{i_4}] \\ &= \sigma^4 \sum_{i_1,i_2,i_3,i_4} \Delta_{i_1i_2i_3i_4}\mathbb{E}[\mathbf{b}_{i_1}\mathbf{b}_{i_2}\mathbf{b}_{i_3}\mathbf{b}_{i_4}] \\ &= \sigma^4\,\mathbb{E}\left[\sum_{i_1,i_3}\mathbf{b}_{i_1}^2\mathbf{b}_{i_3}^2 + \sum_{i_1,i_4}\mathbf{b}_{i_1}^2\mathbf{b}_{i_4}^2 + \sum_{i_1,i_2}\mathbf{b}_{i_1}^2\mathbf{b}_{i_2}^2\right] \\ &= 3\sigma^4\mathbb{E}\,\|\mathbf{B}\|_F^4 .\end{aligned}$$

□

Lemma 4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a random matrix whose entries are i.i.d Gaussian random variables with mean zero and variance σ^2, and let $\mathbf{B} \in \mathbb{R}^{p\times m}$ be a (random) matrix independent of $\mathbf{A}$. Then,*

$$\begin{aligned}\mathbb{E}\,\|\mathbf{BA}\|_F^4 &= n\sigma^4\left(n\,\mathbb{E}\,\|\mathbf{B}\|_F^4 + 2\mathbb{E}\,\mathrm{tr}((\mathbf{B}^\mathsf{T}\mathbf{B})^2)\right) \\ &\leq \sigma^4 n(n+2)\mathbb{E}\,\|\mathbf{B}\|_F^4 .\end{aligned}$$

Proof. By definition of the Frobenius norm we have

$$\mathbb{E}\,\|\mathbf{BA}\|_F^4 = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{BAA}^\mathsf{T}\right)\mathrm{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{BAA}^\mathsf{T}\right)\right].$$

Since $\mathbf{A}_{ij} \sim \mathcal{N}(0,\sigma^2)$ for any $i \in [m], j \in [n]$, $\mathbf{AA}^\mathsf{T} \in \mathbb{R}^{m\times m}$ is a random symmetric positive definite matrix following a Wishart distribution with parameters m, n and $\sigma^2\mathbf{I}_m \in \mathbb{R}^{m\times m}$. Therefore,

$$\begin{aligned}&\mathbb{E}\left[\mathrm{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{BAA}^\mathsf{T}\right)\mathrm{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{BAA}^\mathsf{T}\right)\right] \\ &\quad= n\sigma^4\left(n\mathbb{E}\,\|\mathbf{B}\|_F^4 + 2\mathbb{E}\,\mathrm{tr}((\mathbf{B}^\mathsf{T}\mathbf{B})^2)\right) \\ &\quad\leq \sigma^4 n(n+2)\mathbb{E}\,\|\mathbf{B}\|_F^4 ,\end{aligned}$$

where the equality follows from standard properties of the Wishart distribution (see *e.g.*, Section 3.3.6 of [Gupta and Nagar, 2018]), and the inequality follows from the sub-multiplicativity of the Frobenius norm. □

Let us now start by defining the tensor $\boldsymbol{\mathcal{M}}^n \in \mathbb{R}^{R\times d_1\times d_2\times\cdots\times d_{n-1}}$ for each $2 \leq n \leq N$ component-wise by

$$\boldsymbol{\mathcal{M}}^n_{r,i_1,\ldots,i_{n-1}} = \sum_{\substack{i_n,\ldots,i_N \\ r_n,\ldots,r_{N-1}}} (\boldsymbol{\mathcal{G}}^n)_{r,i_n,r_n}(\boldsymbol{\mathcal{G}}^{n+1})_{r_n,i_{n+1},r_{n+1}} \ldots (\boldsymbol{\mathcal{G}}^{N-1})_{r_{N-2},i_{N-1},r_{N-1}}(\boldsymbol{\mathcal{G}}^N)_{r_{N-1},i_N}\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_N},$$

for each $r \in [R]$, $i_1 \in [d_1], \ldots, i_{n-1} \in [d_{n-1}]$. In some sense, $\boldsymbol{\mathcal{M}}^n$ is the tensor obtained by removing the first $n-1$ cores from the computation of $y_1 = \langle\langle\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^N\rangle\rangle, \boldsymbol{\mathcal{X}}\rangle$. With this definition, one can check that $\bullet\ \langle\langle\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^N\rangle\rangle\boldsymbol{\mathcal{X}}\rangle = \langle \boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{M}}^2\rangle$, $\bullet\ \boldsymbol{\mathcal{M}}^N_{(1)} = (\boldsymbol{\mathcal{G}}^N)_{(1)}\boldsymbol{\mathcal{X}}_{(N)}$ and $\bullet\ \boldsymbol{\mathcal{M}}^n_{(1)} = (\boldsymbol{\mathcal{G}}^n)_{(1)}(\boldsymbol{\mathcal{M}}^{n+1})_{(1,n+1)}$ for each $n \in [N]$, where $(\boldsymbol{\mathcal{M}}^{n+1})_{(1,n+1)} \in \mathbb{R}^{Rd_n\times d_1\ldots d_{n-1}}$ denotes the matricization of $\boldsymbol{\mathcal{M}}^{n+1}$ obtained by mapping its first and last modes to rows and the other ones to columns. Let σ_n^2 denote the variance used to draw the entries of each core $\boldsymbol{\mathcal{G}}^n$. Using Lemma 3 we obtain

$$\begin{aligned}\mathbb{E}y_1^4 &= \mathbb{E}\langle\langle\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^N\rangle\rangle, \boldsymbol{\mathcal{X}}\rangle^4 = \mathbb{E}\langle \boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{M}}^2_{(1)}\rangle^4 \\ &= 3\sigma_1^4\mathbb{E}\left\|\boldsymbol{\mathcal{M}}^2_{(1)}\right\|_F^4 = 3\sigma_1^4\mathbb{E}\left\|(\boldsymbol{\mathcal{G}}^2)_{(1)}\boldsymbol{\mathcal{M}}^3_{(1,3)}\right\|_F^4 .\end{aligned}$$

Using the fact that the Frobenius norm of a tensor is constant across all matricizations and by Lemma 4 we get

$$\begin{aligned}\mathbb{E}[y_1^4] &= 3\sigma_1^4\mathbb{E}\left\|(\boldsymbol{\mathcal{G}}^2)_{(1)}\boldsymbol{\mathcal{M}}^3_{(1,3)}\right\|_F^4 \\ &\leq 3\sigma_1^4\sigma_2^4 R(R+2)\mathbb{E}\left\|\boldsymbol{\mathcal{M}}^3_{(1,3)}\right\|_F^4 \\ &= 3\sigma_1^4\sigma_2^4 R(R+2)\mathbb{E}\left\|\boldsymbol{\mathcal{M}}^3_{(1)}\right\|_F^4 \\ &= 3\sigma_1^4\sigma_2^4 R(R+2)\mathbb{E}\left\|(\boldsymbol{\mathcal{G}}^3)_{(1)}\boldsymbol{\mathcal{M}}^4_{(1,4)}\right\|_F^4 \\ &\leq 3\sigma_1^4\sigma_2^4\sigma_3^4 R^2(R+2)^2\mathbb{E}\left\|\boldsymbol{\mathcal{M}}^4_{(1,4)}\right\|_F^4\end{aligned}$$

Similarly, using successive applications of Lemma 4 it then follows that

$$\begin{aligned}
\mathbb{E}[y_1^4] &\leq 3\sigma_1^4 \cdots \sigma_{N-1}^4 R^{N-2}(R+2)^{N-2}\mathbb{E}\left\|\boldsymbol{\mathcal{M}}_{(1)}^N\right\|_F^4 \\
&= 3\sigma_1^4 \cdots \sigma_{N-1}^4 R^{N-2}(R+2)^{N-2}\mathbb{E}\left\|(\boldsymbol{\mathcal{G}}^N)_{(1)}\boldsymbol{\mathcal{X}}_{(N)}\right\|_F^4 \\
&\leq 3\sigma_1^4 \cdots \sigma_N^4 R^{N-1}(R+2)^{N-1}\left\|\boldsymbol{\mathcal{X}}_{(N)}\right\|_F^4 \\
&= 3\frac{1}{R}\left(\frac{1}{R^2}\right)^{N-2}\frac{1}{R}R^{N-1}(R+2)^{N-1}\left\|\boldsymbol{\mathcal{X}}\right\|_F^4 \\
&= 3\left(1+\frac{2}{R}\right)^{N-1}\left\|\boldsymbol{\mathcal{X}}\right\|_F^4.
\end{aligned}$$

Therefore we obtain $\mathbb{E}\left\|\mathbf{y}\right\|_2^4 \leq 3k\left(1+\frac{2}{R}\right)^{N-1}\left\|\boldsymbol{\mathcal{X}}\right\|_F^4 + k(k-1)\left\|\boldsymbol{\mathcal{X}}\right\|_F^4$. Finally,

$$\begin{aligned}
\mathrm{Var}\left(\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2\right) &= \mathbb{E}[\|k^{-\frac{1}{2}}\mathbf{y}\|_2^4] - \mathbb{E}[\|k^{-\frac{1}{2}}\mathbf{y}\|_2^2]^2 \\
&= \frac{1}{k^2}\mathbb{E}\left\|\mathbf{y}\right\|_2^4 - \left\|\boldsymbol{\mathcal{X}}\right\|_F^4 \\
&\leq \frac{1}{k}\left[3\left(1+\frac{2}{R}\right)^{N-1} - 1\right]\left\|\boldsymbol{\mathcal{X}}\right\|_F^4.
\end{aligned}$$

5.2 Proof of Theorem 2: TT case

Theorem 2 for the map $f_{\mathrm{TT}(R)}$ directly follows from the following concentration bound.

Theorem 5. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$. There exist absolute constants C and $K > 0$ such that the random projection map $f_{\mathrm{TT}(R)}$ (see Definition 1) satisfies*

$$\mathbb{P}\left(\left|\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 - \left\|\boldsymbol{\mathcal{X}}\right\|_F^2\right| \geq \varepsilon\left\|\boldsymbol{\mathcal{X}}\right\|_F^2\right) \leq C\exp\left[-\frac{(\sqrt{k}\varepsilon)^{\frac{1}{N}}}{(3K)^{\frac{1}{2N}}\sqrt{1+2/R}}\right].$$

To show this concentration bound, we will use the following extension of the Hanson-Wright inequality whose proof can be found in [Schudy and Sviridenko, 2012].

Theorem 6. ***(Hypercontractivity Concentration Inequality)*** *Consider a degree q polynomial $f(Y) = f(Y_1, \ldots, Y_n)$ of independent centered Gaussian or Rademacher random variables $Y_1, \ldots, Y_n$. Then for any $\lambda > 0$*

$$\mathbb{P}[|f(Y) - \mathbb{E}\left[f(Y)\right]| \geq \lambda] \leq e^2.e^{-\left(\frac{\lambda^2}{K.\mathrm{Var}([f(Y)])}\right)^{\frac{1}{q}}},$$

where $\mathrm{Var}\left([f(Y)]\right)$ *is the variance of the random variable $f(Y)$ and $K > 0$ is an absolute constant.*

Using the bound on the variance of $\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2$ and the fact that $\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2$ is a polynomial of degree $2N$ of independent Gaussian random variables (the entries of the core tensors $\boldsymbol{\mathcal{G}}_i^1, \boldsymbol{\mathcal{G}}_i^2, \cdots, \boldsymbol{\mathcal{G}}_i^N$), we can use Theorem 6 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\left|\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 - \left\|\boldsymbol{\mathcal{X}}\right\|_F^2\right| \geq \lambda\right] \\
&\leq e^2\exp\left[-\left(\frac{\lambda^2}{K\mathrm{Var}\left(\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2\right)}\right)^{\frac{1}{2N}}\right].
\end{aligned}$$

Let $C = e^2$ and let $\lambda = \varepsilon\left\|\boldsymbol{\mathcal{X}}\right\|_F^2$, we finally get

$$\begin{aligned}
&\mathbb{P}\left[\left|\left\|f_{\mathrm{TT}(R)}(\boldsymbol{\mathcal{X}})\right\|_2^2 - \left\|\boldsymbol{\mathcal{X}}\right\|_F^2\right| \geq \varepsilon\left\|\boldsymbol{\mathcal{X}}\right\|_F^2\right] \\
&\leq C\exp\left[-\left(\frac{k\varepsilon^2\left\|\boldsymbol{\mathcal{X}}\right\|_F^4}{3K(1+2/R)^{N-1}\left\|\boldsymbol{\mathcal{X}}\right\|_F^4}\right)^{\frac{1}{2N}}\right] \\
&\leq C\exp\left[-\frac{(\sqrt{k}\varepsilon)^{\frac{1}{N}}}{(3K)^{\frac{1}{2N}}\sqrt{1+2/R}}\right],
\end{aligned}$$

where the last inequality follows from the fact that

$$(1+2/R)^{\frac{N-1}{2N}} \leq \sqrt{1+2/R}.$$

6 Experiments

In this section we compare the embedding quality of the tensorized projection maps $f_{\mathrm{TT}(R)}$, $f_{\mathrm{CP}(R)}$ and Gaussian RP in a simulation study[§]. In particular, we investigate the effect of the rank parameter R for different sizes and orders of input tensors. We first randomly generate an N-th order d-dimensional tensor $\boldsymbol{\mathcal{X}}$ (*i.e.* vector of size d^N) with unit norm in the TT format with rank $\tilde{R} = 10$. To assess how well the tensorized maps scale to very high order tensors, we consider three cases: • small-order: $(d = 15, N = 3)$, • medium-order: $(d = 3, N = 12)$ and • high-order $(d = 3, N = 25)$.

We compare several values of the rank parameter for the two tensorized map: $R = 4, 25, 100$ for $f_{\mathrm{CP}(R)}$ and $R = 2, 5, 10$ for $f_{\mathrm{TT}(R)}$. Note that these values correspond to roughly the same number of parameters for the two maps since $f_{\mathrm{TT}(R)}$ requires the storage of $(N-2)dR^2 + 2dR$ parameters while $f_{\mathrm{CP}(R)}$ only needs NdR. Additional experiment on image data from the CIFAR-10 dataset [Krizhevsky and Hinton, 2009] are presented in Appendix B.1.

The quality of embedding is evaluated using the distortion ratio metric defined by $D(f, \boldsymbol{\mathcal{X}}) = \left|\frac{\|f(\boldsymbol{\mathcal{X}})\|^2}{\|\boldsymbol{\mathcal{X}}\|^2} - 1\right|$.

[§]For these experiments we use *Tensor Toolbox v3.1* [Bader et al., 2019] and *TT-Toolbox v2.2* [Oseledets et al., 2014].

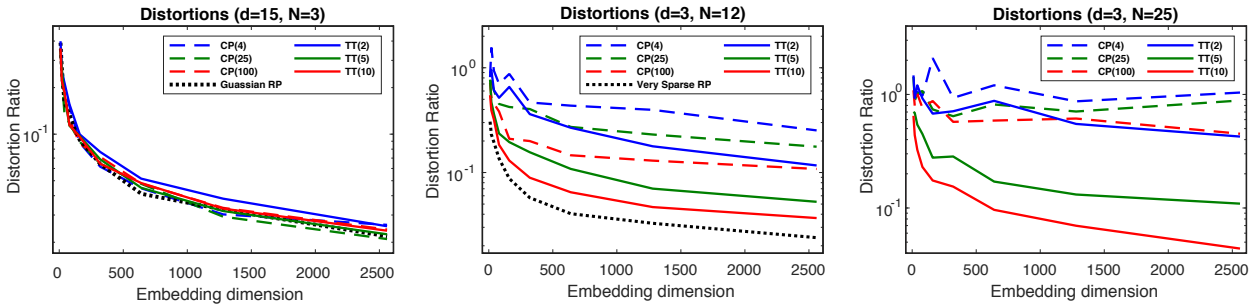

Figure 1: Comparison of the distortion ratio of $f_{\mathrm{TT}(R)}$,$f_{\mathrm{CP}(R)}$, and Gaussian RP for different value of the rank parameter R for small-order (left), medium-order (center) and high-order (right) input tensors.

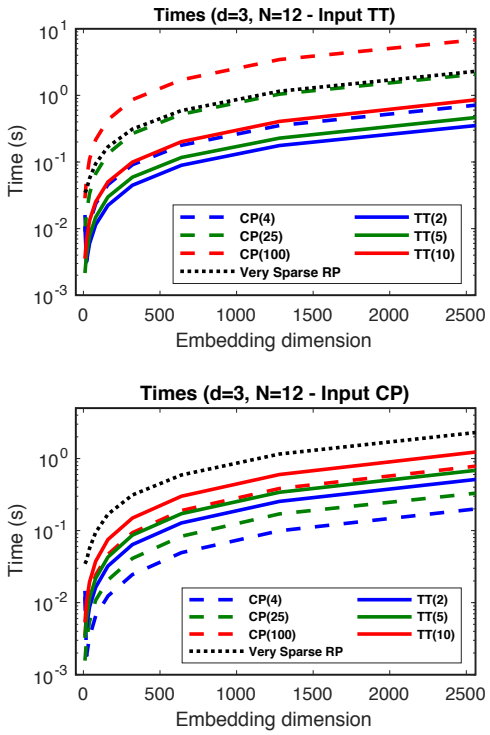

Figure 2: Comparison of embedding time between tensorized and very sparse RP for the medium-order case ($d = 3, N = 12$) when the input is given in the TT format (top) or CP format (bottom).

Due to memory limitation, we compare tensorized RP with Gaussian RP for the small-order case tensors and with very sparse RP [Li et al., 2006] for medium-order tensors (the high-order case cannot be handled with Gaussian or very sparse RP).

The average distortion ratios over 100 trials are reported as a function of the embedding dimension k in Figure 1. In the small-order case, we see that $f_{\mathrm{TT}(R)}$ and $f_{\mathrm{CP}(R)}$ perform similarly to Gaussian RP for all values of the rank parameter. In the medium-order case, we see that the rank of the tensorized RP significantly affects the quality of the embedding. Moreover, $f_{\mathrm{CP}(R)}$ struggles to achieve a good distortion ratio even when $R = 100$ while $f_{\mathrm{TT}(R)}$ almost reaches the performance of very sparse RP. This behavior is accentuated in the high-order case where $f_{\mathrm{CP}(R)}$ obtains poor performances even for high values of R and k while $f_{\mathrm{TT}(R)}$ provides good embeddings for $R = 5, 10$. Note that this behavior is expected from our theoretical analysis.

To illustrate the time complexity of the algorithms, we report the average running time needed to project the input tensor for the medium-order case in Figure 2, when $\boldsymbol{\mathcal{X}}$ is either given as a TT or a CP tensor of rank 10. We see that $f_{\mathrm{TT}(R)}$ (resp. $f_{\mathrm{CP}(R)}$) is more efficient when the input tensor is given in the TT format (resp. CP format), which is somehow expected. We also report the average running time needed to project the different input tensor in medium-order case ($d = 3, N = 8, 11, 12, 13$) with respect to the dimension d^N (Appendix B.2). It is also worth observing that $f_{\mathrm{TT}(R)}$ is always faster than very sparse RP while it is not the case for $f_{\mathrm{CP}(R)}$.

7 Conclusion

We propose a novel efficient RP technique for high-order tensor data: tensorized random projections maps. We theoretically and empirically studied two tensorized maps relying on the CP and TT deocmposition format, respectively. Our theoretical analysis and simulation study show that the TT format is better suited than the CP format for tensorizing random projections.

Future work include leveraging and extending our theoretical results to design efficient sketching algorithms for high-order tensor data. In particular, we plan to develop fast low rank approximation algorithms for matrices given in the TT format, which could prove particularly useful for designing efficient PCA and CCA algorithms for high-dimensional tensor data.

Acknowledgment

This research is supported by the Canadian Institute for Advanced Research (CIFAR AI chair program). This work was completed while Beheshteh T. Rakhshan interned at Montreal Institute for Learning Algorithms (Mila), Montreal, QC.

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. pages 141–160, 2020.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms (TALG)*, 9(3):21, 2013.

Brett W. Bader, Tamara G. Kolda, et al. Matlab tensor toolbox version 3.1. Available online, June 2019. URL https://www.tensortoolbox.org.

Mohammad Taha Bahadori, Qi Rose Yu, and Yan Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in Neural Information Processing Systems*, pages 3491–3499, 2014.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

A. Cichocki, R. Zdunek, A.H. Phan, and S.I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley, 2009.

Sanjoy Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

Sanjoy Dasgupta. Experiments with random projection. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 143–151, 2000.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Ruhui Jin, Tamara G Kolda, and Rachel Ward. Faster Johnson-Lindenstrauss transforms via Kronecker products. *arXiv preprint arXiv:1909.04801*, 2019.

William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1WRibb0Z.

Jon M Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, volume 97, pages 599–608, 1997.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296. ACM, 2006.

H. Lu, K.N. Plataniotis, and A. Venetsanopoulos. *Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data*. CRC Press, 2013.

Alexander Novikov, Anton Rodomanov, Anton Osokin, and Dmitry Vetrov. Putting MRFs on a tensor train. In *International Conference on Machine Learning*, pages 811–819, 2014.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. In *Advances in neural information processing systems*, pages 442–450, 2015.

Ivan Oseledets, Vladimir Kazeev, et al. Matlab tt-toolbox 2.2: Fast multidimensional array operations in tt-format. Available online, June 2014. URL https://github.com/oseledets/TT-Toolbox.

Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for support vector machines. In *In Proceeding of the Artificial Intelligence and Statistics*, pages 498–506, 2013.

Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013.

Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. Society for Industrial and Applied Mathematics, 2012.

Yang Shi and Animashree Anandkumar. Multi-dimensional tensor sketch. *arXiv preprint arXiv:1901.11261*, 2019.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Yiming Sun, Yang Guo, Joel A Tropp, and Madeleine Udell. Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*, 2018.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

H. Zhou, L. Li, and H. Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.