Rademacher Random Projections with Tensor Networks

Beheshteh T. Rakhshan
DIRO & Mila
Université de Montréal
rakhshab@mila.quebec

Guillaume Rabusseau
DIRO & Mila, CIFAR AI Chair
Université de Montréal
grabus@iro.umontreal.ca

Abstract

Random projection (RP) have recently emerged as popular techniques in the machine learning community for their ability in reducing the dimension of very high-dimensional tensors. Following the work in [29], we consider a tensorized random projection relying on Tensor Train (TT) decomposition where each element of the core tensors is drawn from a Rademacher distribution. Our theoretical results reveal that the Gaussian low-rank tensor represented in compressed form in TT format in [29] can be replaced by a TT tensor with core elements drawn from a Rademacher distribution with the same embedding size. Experiments on synthetic data demonstrate that tensorized Rademacher RP can outperform the tensorized Gaussian RP studied in [29]. In addition, we show both theoretically and experimentally, that the tensorized RP in the Matrix Product Operator (MPO) format proposed in [5] for performing SVD on large matrices is not a Johnson-Lindenstrauss transform (JLT) and therefore not a well-suited random projection map.

1 Introduction

Tensor decompositions are popular techniques used to effectively deal with high-dimensional tensor computations. They recently become popular in the machine learning community for their ability to perform operations on very high-order tensors and successfully have been applied in neural networks [23, 24], supervised learning [33, 25], unsupervised learning [32, 22, 11], neuro-imaging [37], computer vision [20] and signal processing [7, 31] to name a few. There are different ways of decomposing high-dimensional tensors efficiently. Two most powerful decompositions, CP [13] and Tucker [35] decompositions, can represent very high-dimensional tensors in a compressed form. However, the number of parameters in the Tucker decomposition grows exponentially with the order of a tensor. While in the CP decomposition, the number of parameters scales better, even computing the rank is an NP-hard problem [12, 17]. Tensor Train (TT) decomposition [28] fixed these challenges as the number of parameters grows linearly with the order of a tensor and enjoys efficient and stable numerical algorithms.

In parallel, recent advances in Random Projections (RPs) and Johnson-Lindestrauss (JL) embeddings have succeeded in scaling up classical algorithms to high-dimensional data [36, 6]. While many efficient random projection techniques have been proposed to deal with high-dimensional vector data [2, 3, 4], it is not the case for high-order tensors. To address this challenge, it is crucial to find efficient RPs to deal with the curse of dimensionality caused by very high-dimensional data. Recent advances in employing JL transforms for dealing with high-dimensional tensor inputs offer efficient embeddings for reducing computational costs and memory requirements [29, 15, 34, 21, 18, 8]. In particular, Batselier et al. [5] used the Matrix Product Operator (MPO) format to propose an algorithm (TNrSVD) for randomized SVD of very high-dimensional matrices. At the same time, [29]

Second Workshop on Quantum Tensor Networks in Machine Learning. In conjunction with 35th NeurIPS, 2021

used TT decomposition to speed up classical Gaussian RP for very high-dimensional input tensors efficiently without flattening the structure of the input into a vector.

Our contribution is two-fold. First, we show that tensorizing an RP using the MPO format does not lead to a JL transform by showing that even in the case of matrix inputs, the variance of such a map does not decrease to zero as the size of embedding increases. Second, our results demonstrate that the tensorized Gaussian RP in [29] can be replaced by a simpler and faster projection using a Rademacher distribution instead of a standard Gaussian distribution. We propose a tensorized RP akin to tensorized Gaussian RP by enforcing each row of a matrix $\mathbf{A} \in \mathbb{R}^{k \times d^N}$ where $k \ll d^N$ to have a low rank tensor structure (TT decomposition) with core elements drawn independently from a Rademacher distribution. Our results show that the Rademacher projection map still benefits from JL transform properties while preserving the same bounds as the tensorized Gaussian RP without any sacrifice in quality of the embedding size. Experiments show that in practice, the performance of the tensorized RP with Rademacher random variables outperforms tensorized Gaussian RP since it reduces the number of operations as it does not require any multiplication.

2 Preliminaries

Lower case bold letters denote vectors, *e.g.* $\mathbf{a}$, upper case bold letters denote matrices, *e.g.* $\mathbf{A}$, and bold calligraphic letters denote higher order tensors, *e.g.* $\boldsymbol{\mathcal{A}}$. The 2-norm of a vector $\mathbf{v}$ is denoted by $\|\mathbf{v}\|_2$ or simply $\|\mathbf{v}\|$. The symbol "$\circ$" denotes the outer product (or tensor product) between vectors. We use $\text{vec}(\mathbf{M}) \in \mathbb{R}^{d_1 \cdot d_2}$ to denote the column vector obtained by concatenating the columns of the matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$. The $d \times d$ identity matrix is denoted by $\mathbf{I}_d$. For any integer i we use $[i]$ to denote the set of integers from 1 to i.

2.1 Tensors

A tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is a multidimensional array and its Frobenius norm is defined by $\|\boldsymbol{\mathcal{T}}\|_F^2 = \langle \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{T}} \rangle$. 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. Let $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ be an N-way tensor. For $n \in [N]$, let $\boldsymbol{\mathcal{G}}^n \in \mathbb{R}^{R_{n-1} \times d_n \times R_n}$ be 3rd order core tensors with $R_0 = R_N = 1$ and $R_1 = \cdots = R_{N-1} = R$. A rank R *tensor train decomposition* of $\boldsymbol{\mathcal{S}}$ is given 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.

Suppose $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{I_1 \times J_1 \times \cdots \times I_N \times J_N}$. For $n \in [N]$, let $\boldsymbol{\mathcal{A}}^n \in \mathbb{R}^{R_{n-1} \times I_n \times J_n \times R_n}$ with $R_0 = R_N = 1$ and $R_1 = \cdots = R_{N-1} = R$. A rank R MPO decomposition of $\boldsymbol{\mathcal{T}}$ is given by $\boldsymbol{\mathcal{T}}_{i_1,j_1,\ldots,i_N,j_N} = (\boldsymbol{\mathcal{A}}^1)_{i_1,j_1,:}(\boldsymbol{\mathcal{A}}^2)_{:,i_2,j_2,:} \ldots (\boldsymbol{\mathcal{A}}^{N-1})_{:,i_{N-1},j_{N-1},:}(\boldsymbol{\mathcal{A}}^N)_{:,i_N,j_N}$ for all indices $i_1 \in [I_1], \cdots, i_N \in [I_N]$ and $j_1 \in [J_1], \ldots, j_N \in [J_N]$; we will use the notation $\boldsymbol{\mathcal{T}} = \text{MPO}((\boldsymbol{\mathcal{A}}^n)_{n=1}^N)$ to denote the MPO format.

2.2 Random Projection

Random projections (RP) are efficient tools for projecting linearly high-dimensional data down into a lower dimensional space while preserving the pairwise distances between points. This is the classical result of the Johnson-Lindenstrauss lemma [16] which states that any n-point set $P \subseteq \mathbb{R}^d$ can be projected linearly into a k-dimensional space with $k = \Omega(\varepsilon^{-2} \log(n))$. One of the simplest way to generate such a projection is using a $d \times k$ random Gaussian matrix $\mathbf{A}$, i.e., the entries of $\mathbf{A}$ are drawn independently from a standard Gaussian distribution with mean zero and variance one. More precisely, for any two points $\mathbf{u}, \mathbf{v} \in P \subseteq \mathbb{R}^d$ the following inequality holds with high probability

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

where $f : \mathbb{R}^d \to \mathbb{R}^k$ ($k \ll d$) is a linear map $f(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}$ and $\mathbf{A} \in \mathbb{R}^{k \times d}$ is a random matrix. We also call f a Johnson-Lindenstrauss transform (JLT). To have a JLT, the random projection map f must satisfy the following two properties: (i) Expected isometry, i.e., $\mathbb{E}\left[\|f(\mathbf{x})\|^2\right] = \|\mathbf{x}\|^2$ and (ii) Vanishing variance, that is Var $\left(\|f(\mathbf{x})\|^2\right)$ decreases to zero as the embedding size k increases.

3 Random Projections based on Tensor Decomposition

3.1 Matrix Product Operator Random Projection

Classical random projection maps $f : \mathbf{x} \rightarrow \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}$ deal with high-dimensional data using a dense random matrix $\mathbf{A}$. Due to storage and computational constraints, sparse and very sparse RPs have been proposed in [1, 19], but even sparse RPs still suffer from the curse of dimensionality and cannot handle high-dimensional tensor inputs. To alleviate this difficulty, tensor techniques can be used to compress RP maps. One natural way for this purpose is to compress the dense matrix $\mathbf{A}$ with the Matrix Product Operator (MPO) format [27]. Relying on the MPO format, we can define a random projection map which embeds any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times\cdots\times d_N}$ into $\mathbb{R}^k$, where $k \ll d_1 d_2 \cdots d_N$ is the embedding dimension and is defied element-wise by

$$f(\boldsymbol{\mathcal{X}})_k = \frac{1}{\sqrt{R^{N-1}k}} \sum_{i_1,\dots i_N} \big(\mathrm{MPO}((\boldsymbol{\mathcal{G}}^n)_{n=1}^N)\big)_{i_1,\dots,i_N,k} \boldsymbol{\mathcal{X}}_{i_1,\dots i_N}, \quad i_n \in [d_n] \text{ for } 1 \le n \le N \tag{1}$$

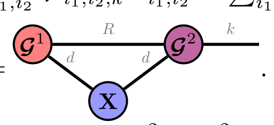
where $\boldsymbol{\mathcal{G}}^1 \in \mathbb{R}^{1\times d_1 \times 1 \times R}, \boldsymbol{\mathcal{G}}^2 \in \mathbb{R}^{R\times d_2\times 1\times R}, \cdots, \boldsymbol{\mathcal{G}}^{N-1} \in \mathbb{R}^{R\times d_{N-1}\times 1\times R}, \boldsymbol{\mathcal{G}}^N \in \mathbb{R}^{R\times d_N\times k\times 1}$ and the entries of each $\boldsymbol{\mathcal{G}}^n$ for $n \in [N]$ are drawn independently from standard Gaussian distribution. We call the map defined in eqn. 1 an MPO RP.

Even though this map satisfies expected isometry property, it is not JLT as its variance does not decrease to zero when the size of random dimension increases.

Proposition 1. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times\cdots\times d_N}$. The MPO RP defined in eqn.* (1) *satisfies the following properties*

- $\mathbb{E}\,\|f(\boldsymbol{\mathcal{X}})\|_2^2 = \|\boldsymbol{\mathcal{X}}\|_F^2$
- $\mathrm{Var}\left(\|f(\mathbf{X})\|_2^2\right) = \frac{2}{k}\,\|\mathbf{X}\|_F^4 + \frac{2}{R}(1+\frac{2}{k})\,\mathrm{tr}((\mathbf{X}^\top\mathbf{X})^2)$ *for $N = 2$*

Proof. We start by showing the expected isometry property. For a fixed $\kappa \in [k]$, suppose $\mathbf{y}_\kappa = \sum_{i_1,\dots i_N} \left(\mathrm{MPO}((\boldsymbol{\mathcal{G}}^n)_{n=1}^N)\right)_{i_1,\dots,i_N,\kappa} \boldsymbol{\mathcal{X}}_{i_1,\dots i_N}$ and $\boldsymbol{\mathcal{S}}_\kappa = \mathrm{MPO}((\boldsymbol{\mathcal{G}}^n)_{n=1}^N)_{:,:,\dots,:,\kappa}$. With these definitions $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_k]$ and $f(\boldsymbol{\mathcal{X}}) = \frac{1}{\sqrt{R^{N-1}k}}\mathbf{y}$. As it is shown in [29] (e.g., see section 5.1), for $\boldsymbol{\mathcal{T}} = \langle\!\langle \boldsymbol{\mathcal{M}}^1, \dots, \boldsymbol{\mathcal{M}}^N \rangle\!\rangle$ with the entries of each core tenors drawn independently from a Gaussian distribution with mean zero and variance one, we have $\mathbb{E}\langle \boldsymbol{\mathcal{T}} \otimes \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}} \otimes \boldsymbol{\mathcal{X}} \rangle = R^{N-1}\|\boldsymbol{\mathcal{X}}\|_F^2$. Therefore, $\boldsymbol{\mathcal{S}}_\kappa = \langle\!\langle \boldsymbol{\mathcal{G}}^1, \dots, \boldsymbol{\mathcal{G}}^N_{:,:,\kappa} \rangle\!\rangle$ and $\mathbb{E}[\mathbf{y}_\kappa^2] = \mathbb{E}[\langle \boldsymbol{\mathcal{S}}_\kappa \otimes \boldsymbol{\mathcal{S}}_\kappa, \boldsymbol{\mathcal{X}} \otimes \boldsymbol{\mathcal{X}} \rangle] = R^{N-1}\|\boldsymbol{\mathcal{X}}\|_F^2$. From which we can conclude $\mathbb{E}[\|f(\boldsymbol{\mathcal{X}})\|^2] = \frac{1}{R^{N-1}k}\sum_\kappa \mathbb{E}[\mathbf{y}_\kappa^2] = \|\boldsymbol{\mathcal{X}}\|_F^2$.

Now, in order to find a bound for variance of $\|\mathbf{y}\|_2^2$ we need first to find a bound for $\mathbb{E}[\|\mathbf{y}\|_2^4]$. For $N = 2$, let $\boldsymbol{\mathcal{T}} = \mathrm{MPO}(\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2)$ and $\mathbf{y}_k = \sum_{i_1,i_2} \boldsymbol{\mathcal{T}}_{i_1,i_2,k}\mathbf{X}_{i_1,i_2} = \sum_{i_1,i_2}\sum_r \boldsymbol{\mathcal{G}}^1_{i_1 r}\boldsymbol{\mathcal{G}}^2_{r i_2 k}\mathbf{X}_{i_1 i_2}$. In terms of tensor network diagrams, we have $\mathbf{y} =$ 

. By defining a tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{d\times R\times d\times R}$ element-wise via $\boldsymbol{\mathcal{M}}_{i_1 r_1 i_2 r_2} = \sum_{j_1,j_2,k} \mathbf{X}_{i_1 j_1}\boldsymbol{\mathcal{G}}^2_{j_1 r_1 k}\boldsymbol{\mathcal{G}}^2_{j_2 r_2 k}\mathbf{X}_{i_2 j_2}$, since $\boldsymbol{\mathcal{G}}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using Isserlis' theorem [14] gives us

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{y}\|_2^4] = \mathbb{E}[\langle(\boldsymbol{\mathcal{G}}^1)^{\otimes 4}, \boldsymbol{\mathcal{M}}^{\otimes 2}\rangle] = \langle \mathbb{E}[(\boldsymbol{\mathcal{G}}^1)^{\otimes 4}], \mathbb{E}[\boldsymbol{\mathcal{M}}^{\otimes 2}]\rangle \\
&= \sum_{i_1,\dots,i_4}\sum_{r_1,\dots,r_4} \mathbb{E}[\boldsymbol{\mathcal{G}}^1_{i_1 r_1}\boldsymbol{\mathcal{G}}^1_{i_2 r_2}\boldsymbol{\mathcal{G}}^1_{i_3 r_3}\boldsymbol{\mathcal{G}}^1_{i_4 r_4}]\mathbb{E}[\boldsymbol{\mathcal{M}}_{i_1 r_1 i_2 r_2}\boldsymbol{\mathcal{M}}_{i_3 r_3 i_4 r_4}] \\
&= \sum_{\substack{i_1,\dots,i_4 \\ r_1,\cdots r_4}} \mathbb{E}\left(\delta_{i_1 i_2}\delta_{i_3 i_4}\delta_{r_1 r_2}\delta_{r_3 r_4} + \delta_{i_1 i_3}\delta_{i_2 i_4}\delta_{r_1 r_3}\delta_{r_2 r_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\delta_{r_1 r_4}\delta_{r_2 r_3}\right)\mathbb{E}[\boldsymbol{\mathcal{M}}_{i_1 r_1 i_2 r_2}\boldsymbol{\mathcal{M}}_{i_3 r_3 i_4 r_4}] \\
&= \mathbb{E}\sum_{\substack{i_1,i_3 \\ r_1,r_3}} \boldsymbol{\mathcal{M}}_{i_1 r_1 i_1 r_1}\boldsymbol{\mathcal{M}}_{i_3 r_3 i_3 r_3} + \mathbb{E}\sum_{\substack{i_1,i_4 \\ r_1,r_4}} \boldsymbol{\mathcal{M}}_{i_1 r_1 i_4 r_4}\boldsymbol{\mathcal{M}}_{i_1 r_1 i_4 r_4} + \mathbb{E}\sum_{\substack{i_1,i_2 \\ r_1,r_2}} \boldsymbol{\mathcal{M}}_{i_1 r_1 i_2 r_2}\boldsymbol{\mathcal{M}}_{i_2 r_2 i_1 r_1} \\
&= \mathbb{E}\left[\mathrm{tr}\left(\mathbf{X}\boldsymbol{\mathcal{G}}^2_{(2)}(\boldsymbol{\mathcal{G}}^2_{(2)})^\top\mathbf{X}^\top\right)\mathrm{tr}\left(\mathbf{X}\boldsymbol{\mathcal{G}}^2_{(2)}(\boldsymbol{\mathcal{G}}^2_{(2)})^\top\mathbf{X}^\top\right)\right] + 2\mathbb{E}\sum_{i_1,i_4}\sum_{r_1,r_4} \boldsymbol{\mathcal{M}}_{i_1 r_1 i_4 r_4}\boldsymbol{\mathcal{M}}_{i_1 r_1 i_4 r_4},
\end{aligned}$$

where the second term in above equation is obtained by using symmetry property of tensor $\boldsymbol{\mathcal{M}}$, i.e., $\boldsymbol{\mathcal{M}}_{i_1r_1i_2r_2} = \boldsymbol{\mathcal{M}}_{i_2r_2i_1r_1}$. Since $\boldsymbol{\mathcal{G}}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\mathcal{G}}^2_{(2)}(\boldsymbol{\mathcal{G}}^2_{(2)})^\top \in \mathbb{R}^{d\times d}$ is a random symmetric positive definite matrix, by standard properties of the Wishart distribution (see *e.g.*, Section 3.3.6 of [9]) we have $R^2k^2\|\mathbf{X}\|_F^2 + 2Rk\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2) + 2\mathbb{E}\sum_{i_1,i_4}\sum_{r_1,r_4}\boldsymbol{\mathcal{M}}_{i_1r_1i_4r_4}\boldsymbol{\mathcal{M}}_{i_1r_1i_4r_4}$. Again, by using Isserlis' theorem element-wise for the tensor $\boldsymbol{\mathcal{G}}^2$, we can simplify the third term in above equation

$$
\begin{aligned}
&\mathbb{E}\sum_{i_1,i_4}\sum_{r_1,r_4}\boldsymbol{\mathcal{M}}_{i_1r_1i_4r_4}\boldsymbol{\mathcal{M}}_{i_1r_1i_4r_4}\\
&=\mathbb{E}\sum_{i_1,i_4}\sum_{r_1,r_4}\sum_{j_1,j_2,k_1}\sum_{j_3,j_4,k_2}\left(\mathbf{X}_{i_1j_1}\boldsymbol{\mathcal{G}}^2_{j_1r_1k_1}\boldsymbol{\mathcal{G}}^2_{j_2r_4k_1}\mathbf{X}_{i_4j_2}\right)\left(\mathbf{X}_{i_1j_3}\boldsymbol{\mathcal{G}}^2_{j_3r_1k_2}\boldsymbol{\mathcal{G}}^2_{j_4r_4k_2}\mathbf{X}_{i_4j_4}\right)\\
&=\mathbb{E}\sum_{\substack{i_1,i_4\\r_1,r_4}}\sum_{\substack{j_1,j_2,k_1\\j_3,j_4,k_2}}\left(\delta_{j_1j_2}\delta_{j_3j_4}\delta_{r_1r_4}+\delta_{j_1j_3}\delta_{j_2j_4}\delta_{k_1k_2}+\delta_{j_1j_4}\delta_{j_3j_2}\delta_{k_1k_2}\delta_{r_1r_4}\right)\mathbf{X}_{i_1j_1}\mathbf{X}_{i_4j_2}\mathbf{X}_{i_1j_3}\mathbf{X}_{i_4j_4}\\
&=\mathbb{E}\sum_{\substack{i_1,i_4,r_1\\j_1,j_3,k_1,k_2}}\mathbf{X}_{i_1j_1}\mathbf{X}_{i_4j_1}\mathbf{X}_{i_1j_3}\mathbf{X}_{i_4j_3}+\mathbb{E}\sum_{\substack{i_1,i_4,j_1,j_4\\k_1,k_2}}\mathbf{X}_{i_1j_1}\mathbf{X}_{i_4j_4}\mathbf{X}_{i_1j_1}\mathbf{X}_{i_4j_4}+\mathbb{E}\sum_{\substack{i_1,i_4,r_1,r_4\\j_1,j_2k_1,k_2}}\mathbf{X}_{i_1j_1}\mathbf{X}_{i_4j_2}\mathbf{X}_{i_1j_2}\mathbf{X}_{i_4j_1}\\
&=Rk^2\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2)+kR^2\|\mathbf{X}\|_F^4+kR\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2).
\end{aligned}
$$

Therefore,

$$\mathbb{E}[\|\mathbf{y}\|_2^4] = R^2k(k+2)\|\mathbf{X}\|_F^2 + 2kR(2+k)\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2).$$

Finally,

$$
\begin{aligned}
\operatorname{Var}\left(\|f(\mathbf{X})\|_2^2\right) &= \mathbb{E}[\|k^{-\frac{1}{2}}R^{-\frac{1}{2}}\mathbf{y}\|_2^4] - \mathbb{E}[\|k^{-\frac{1}{2}}R^{-\frac{1}{2}}\mathbf{y}\|_2^2]^2 = \frac{1}{k^2R^2}\mathbb{E}\|\mathbf{y}\|_2^4 - \|\mathbf{X}\|_F^4\\
&= \frac{1}{k^2R^2}\left(R^2k(k+2)\|\mathbf{X}\|_F^2 + 2kR(2+k)\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2)\right) - \|\mathbf{X}\|_F^4\\
&= \frac{2}{k}\|\mathbf{X}\|_F^4 + \frac{2}{R}(1+\frac{2}{k})\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^2). \qquad \square
\end{aligned}
$$

As we can see for $N = 2$, by increasing k the variance does not vanish which validates the fact that the map in eqn. (1) is not a JLT. Using the MPO format to perform a randomized SVD for larges matrices was proposed in [5] for the first time. As mentioned by the authors, even though numerical experiments demonstrate promising results, the paper suffers from a lack of theoretical guarantees (e.g., such as probabilistic bounds for the classical randomized SVD [10]). The result we just showed in Proposition 1 actually demonstrates that obtaining such guarantees is not possible, since the underlying MPO RP used in [5] is not a JLT. As shown in [29] this problem can be fixed by enforcing a low rank tensor structure on the rows of the random projection matrix.

3.2 Tensor Train Random Projection with Rademacher Variables

We now formally define the map proposed by Rakhshan and Rabusseau and show that the probabilistic bounds obtained in [29] can be extended to the Rademacher distribution.

Following the lines in the work done by [29] and due to the computational efficiency of TT decomposition, we propose a similar map to $f_{\mathrm{TT}(R)}$ by enforcing a low rank TT structure on the rows of $\mathbf{A}$, where for each row of $\mathbf{A}$ the core elements are drawn independently from $\{-1, 1\}$ with probability $1/2$, i.e., Rademacher distribution. We generalize and unify the definition of $f_{\mathrm{TT}(R)}$ with Rademacher random projection by first defining the TT distribution and then TT random projection.

Definition 1. *A tensor $\boldsymbol{\mathcal{T}}$ is drawn from a* TT-Gaussian (resp. TT-Rademacher) distribution *with rank parameter R, denoted by $\boldsymbol{\mathcal{T}} \sim \mathrm{TT}_{\mathcal{N}}(R)$ (resp. $\boldsymbol{\mathcal{T}} \sim \mathrm{TT}_{Rad}(R)$), if*

$$\boldsymbol{\mathcal{T}} = \frac{1}{\sqrt{R^{(N-1)}}}\langle\langle\boldsymbol{\mathcal{G}}^1, \boldsymbol{\mathcal{G}}^2, \cdots, \boldsymbol{\mathcal{G}}^N\rangle\rangle,$$

where $\boldsymbol{\mathcal{G}}^1 \in \mathbb{R}^{1\times d_1\times R}, \boldsymbol{\mathcal{G}}^2 \in \mathbb{R}^{R\times d_2\times R}, \cdots, \boldsymbol{\mathcal{G}}^{N-1} \in \mathbb{R}^{R\times d_{N-1}\times R}, \boldsymbol{\mathcal{G}}^N \in \mathbb{R}^{R\times d_N\times 1}$ and the entries of each $\boldsymbol{\mathcal{G}}^n$ for $n \in [N]$ are drawn independently from the standard normal distribution (resp. the Rademacher distribution).

Definition 2. *A* TT Gaussian (resp. TT Rademacher) 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 component-wise by*

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

where $\boldsymbol{\mathcal{T}}_i \sim \mathrm{TT}_{\mathcal{N}}(R)$ *(resp.* $\boldsymbol{\mathcal{T}}_i \sim \mathrm{TT}_{Rad}(R)$*).*

Our main results show that the tensorized Rademacher random projection still benefits from JLT properties as it is an expected isometric map and the variance decays to zero as the random dimension size grows. The following theorems state that using Rademacher random variables instead of standard Gaussian random variables gives us the same results for the bound of the variance while preserving the same lower bound for the size of the random dimension k.

Theorem 2. *Let* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ *and let* $f_{\mathrm{TT}(R)}$ *be either a tensorized Gaussian RP or a tensorized Rademacher RP of rank* R *(see Definition 2) . The random projection map* $f_{\mathrm{TT}(R)}$ *satisfies the following properties:*

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

Proof. The proof for the Gaussian TT random projection is given in [29]. We now show the result for the tensorized Rademacher RP. The proof of the expected isometry part follows the exact same technique as in [29] (see section 5.1, expected isometry part), we thus omit it here. Our proof to bound the variance of $f_{\mathrm{TT}(R)}$ when the core elements are drawn independently from a Rademacher distribution relies on the following lemmas.

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

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

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

$$\mathbb{E}\langle \mathbf{A}, \mathbf{B} \rangle^4 = \mathbb{E}\langle \mathbf{a}, \mathbf{b} \rangle^4 = \mathbb{E}\langle \mathbf{a}^{\otimes 4}, \mathbf{b}^{\otimes 4} \rangle = \sum_{i_1, i_2, i_3, i_4} \mathbb{E}[\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \mathbf{a}_{i_3}, \mathbf{a}_{i_4}] \mathbb{E}[\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \mathbf{b}_{i_3}, \mathbf{b}_{i_4}],$$

we can see that in four cases we have non-zero values for $\mathbb{E}[\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \mathbf{a}_{i_3}, \mathbf{a}_{i_4}]$, i.e.,

$$\mathbb{E}[\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \mathbf{a}_{i_3}, \mathbf{a}_{i_4}] = \begin{cases} 1 & \text{if } i_1 = i_2 = i_3 = i_4 \ \text{ or} \\ & i_1 = i_2 \neq i_3 = i_4 \ \text{ or} \\ & i_1 = i_3 \neq i_2 = i_4 \ \text{ or} \\ & i_1 = i_4 \neq i_2 = i_3. \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Therefore,

$$\mathbb{E}\langle \mathbf{A}, \mathbf{B} \rangle^4 = \sum_{i_1} \mathbb{E}[\mathbf{a}_{i_1}^4] \mathbb{E}[\mathbf{b}_{i_1}^4] + \sum_{i_1 \neq i_3} \mathbb{E}[\mathbf{a}_{i_1}^2] \mathbb{E}[\mathbf{a}_{i_3}^2] \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_3}^2] + \sum_{i_1 \neq i_4} \mathbb{E}[\mathbf{a}_{i_1}^2] \mathbb{E}[\mathbf{a}_{i_4}^2] \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_4}^2]$$
$$+ \sum_{i_1 \neq i_2} \mathbb{E}[\mathbf{a}_{i_1}^2] \mathbb{E}[\mathbf{a}_{i_2}^2] \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_2}^2].$$

Since $\mathbb{E}[\mathbf{a}_{i_1}^4] = \mathbb{E}[\mathbf{a}_{i_1}^2] = \mathbb{E}[\mathbf{a}_{i_2}^2] = \mathbb{E}[\mathbf{a}_{i_3}^2] = \mathbb{E}[\mathbf{a}_{i_4}^2] = 1$, the equation above can be simplified as

$$\begin{aligned} \mathbb{E}\langle \mathbf{A}, \mathbf{B} \rangle^4 &= \sum_{i_1} \mathbb{E}[\mathbf{b}_{i_1}^4] + \sum_{i_1 \neq i_3} \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_3}^2] + \sum_{i_1 \neq i_4} \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_4}^2] + \sum_{i_1 \neq i_2} \mathbb{E}[\mathbf{b}_{i_1}^2] \mathbb{E}[\mathbf{b}_{i_2}^2] \\ &= \mathbb{E} \sum_{i_1} \mathbf{b}_{i_1}^4 + \mathbb{E} \sum_{i_1, i_3} \mathbf{b}_{i_1}^2 \mathbf{b}_{i_3}^2 - \mathbb{E} \sum_{i_1 = i_3} \mathbf{b}_{i_1}^4 + \mathbb{E} \sum_{i_1, i_4} \mathbf{b}_{i_1}^2 \mathbf{b}_{i_4}^2 \\ &- \mathbb{E} \sum_{i_1 = i_4} \mathbf{b}_{i_1}^4 + \mathbb{E} \sum_{i_1, i_2} \mathbf{b}_{i_1}^2 \mathbf{b}_{i_2}^2 - \mathbb{E} \sum_{i_1 = i_2} \mathbf{b}_{i_1}^4 = 3\mathbb{E}\|\mathbf{B}\|_F^4 - 2\mathbb{E}\|\mathbf{b}\|_4^4 \leq 3\mathbb{E}\|\mathbf{B}\|_F^4. \end{aligned}$$

∎

Lemma 4. *Let $\mathbf{A} \in \mathbb{R}^{d\times R}$ be a random matrix whose entries are i.i.d Rademacher random variables with mean zero and variance one, and let $\mathbf{B} \in \mathbb{R}^{p\times d}$ be a random matrix independent of $\mathbf{A}$, then*

$$\mathbb{E}\|\mathbf{BA}\|_F^4 \le R(R+2)\mathbb{E}\|\mathbf{B}\|_F^4.$$

Proof. Setting $\mathbf{M} = \mathbf{B}^\top\mathbf{B}$ we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{BA}\|_F^4 &= \mathbb{E}\left[\operatorname{tr}\left(\mathbf{B}^\top\mathbf{BAA}^\top\right)\operatorname{tr}\left(\mathbf{B}^\top\mathbf{BAA}^\top\right)\right] = \mathbb{E}\langle \mathbf{M}, \mathbf{AA}^\top\rangle^2 \\
&= \sum_{i_1,i_2,i_3,i_4} \mathbb{E}[(\mathbf{AA}^\top)_{i_1,i_2}(\mathbf{AA}^\top)_{i_3,i_4}]\mathbb{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_3i_4}] \\
&= \sum_{i_1,i_2,i_3,i_4}\sum_{j,k}\mathbb{E}[\mathbf{A}_{i_1j}\mathbf{A}_{i_2j}\mathbf{A}_{i_3k}\mathbf{A}_{i_4k}]\mathbb{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_3i_4}].
\end{aligned}$$

Since the components of $\mathbf{A}$ are drawn from a Rademacher distribution, the non-zero summands in the previous equation can be grouped in four cases (which follows from Eq. (2)):

$$\begin{aligned}
\mathbb{E}\|\mathbf{BA}\|_F^4 = &\sum_{i_1\in[d]}\sum_{j,k\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_1k}^2]\mathbb{E}[\mathbf{M}_{i_1i_1}^2] && (i_1 = i_2 = i_3 = i_4) \\
&+ \sum_{\substack{i_1\in[d],\\ i_3\in[d]\setminus\{i_1\}}}\sum_{j,k\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_3k}^2]\mathbb{E}[\mathbf{M}_{i_1i_1}\mathbf{M}_{i_3i_3}] && (i_1 = i_2 \ne i_3 = i_4) \\
&+ \sum_{\substack{i_1\in[d],\\ i_2\in[d]\setminus\{i_1\}}}\sum_{j,k\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}\mathbf{A}_{i_2j}\mathbf{A}_{i_2k}\mathbf{A}_{i_1k}]\mathbb{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_1}] && (i_1 = i_4 \ne i_2 = i_3) \\
&+ \sum_{\substack{i_1\in[d],\\ i_4\in[d]\setminus\{i_1\}}}\sum_{j,k\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}\mathbf{A}_{i_4j}\mathbf{A}_{i_1k}\mathbf{A}_{i_4k}]\mathbb{E}[\mathbf{M}_{i_1i_4}^2] && (i_1 = i_3 \ne i_2 = i_4)
\end{aligned}$$

Now by splitting the summations over $j,k \in [R]$ in two cases $j = k$ and $j \ne k$, and observing that the 3rd and 4th summands in the previous equation vanish when $j \ne k$, we obtain

$$\begin{aligned}
&\mathbb{E}\|\mathbf{BA}\|_F^4 \\
&\quad= \sum_{i_1\in[d]}\sum_{j\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^4]\mathbb{E}[\mathbf{M}_{i_1i_1}^2] + \sum_{i_1\in[d]}\sum_{\substack{j\in[R],\\ k\in[R]\setminus\{j\}}}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_1k}^2]\mathbb{E}[\mathbf{M}_{i_1i_1}^2] \\
&\quad+ \sum_{\substack{i_1\in[d],\\ i_3\in[d]\setminus\{i_1\}}}\sum_{j\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_3j}^2]\mathbb{E}[\mathbf{M}_{i_1i_1}\mathbf{M}_{i_3i_3}] + \sum_{\substack{i_1\in[d],\\ i_3\in[d]\setminus\{i_1\}}}\sum_{\substack{j\in[R],\\ k\in[R]\setminus\{j\}}}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_3k}^2]\mathbb{E}[\mathbf{M}_{i_1i_1}\mathbf{M}_{i_3i_3}] \\
&\quad+ \sum_{\substack{i_1\in[d],\\ i_2\in[d]\setminus\{i_1\}}}\sum_{j\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_2j}^2]\mathbb{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_1}] + \sum_{\substack{i_1\in[d],\\ i_4\in[d]\setminus\{i_1\}}}\sum_{j\in[R]}\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_4j}^2]\mathbb{E}[\mathbf{M}_{i_1i_4}^2].
\end{aligned}$$

Since $\mathbb{E}[\mathbf{A}_{i_1j}^4] = 1$ and $\mathbb{E}[\mathbf{A}_{i_1j}^2\mathbf{A}_{i_1k}^2] = 1$ whenever $j \ne k$, it follows that

$$\begin{aligned}
\mathbb{E}\|\mathbf{BA}\|_F^4 &= R^2\left(\sum_{i_1\in[d]}\mathbb{E}[\mathbf{M}_{i_1i_1}^2] + \sum_{\substack{i_1\in[d],\\ i_3\in[d]\setminus\{i_1\}}}\mathbb{E}[\mathbf{M}_{i_1i_1}\mathbf{M}_{i_3i_3}]\right) + R\left(\sum_{\substack{i_1\in[d],\\ i_2\in[d]\setminus\{i_1\}}}\mathbb{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_1}] + \sum_{\substack{i_1\in[d],\\ i_4\in[d]\setminus\{i_1\}}}\mathbb{E}[\mathbf{M}_{i_1i_4}^2]\right) \\
&= R^2\,\mathbb{E}\sum_{i_1,i_3\in[d]}\mathbf{M}_{i_1i_1}\mathbf{M}_{i_3i_3} + R\,\mathbb{E}\sum_{i_1,i_2\in[d]}\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_1} + R\,\mathbb{E}\sum_{i_1,i_4\in[d]}\mathbf{M}_{i_1i_4}^2 - 2R\,\mathbb{E}\sum_{i_1\in[d]}\mathbf{M}_{i_1i_1}^2 \\
&\le R^2\,\mathbb{E}[\operatorname{tr}(\mathbf{M})^2] + R\,\mathbb{E}\sum_{i_1,i_2\in[d]}\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_1} + R\,\mathbb{E}\sum_{i_1,i_4\in[d]}\mathbf{M}_{i_1i_4}^2 \\
&= R^2\mathbb{E}[\operatorname{tr}(\mathbf{B}^\top\mathbf{B})^2] + 2R\mathbb{E}[\operatorname{tr}((\mathbf{B}^\top\mathbf{B})^2)],
\end{aligned}$$

where in the last equation, we used the fact that $\mathbf{M} = \mathbf{B}^\top\mathbf{B}$ is symmetric. Finally, by the submultiplicavity property of the Frobenius norm, we obtain

$$\mathbb{E}\|\mathbf{BA}\|_F^4 = R^2\mathbb{E}\|\mathbf{B}\|_F^4 + 2R\mathbb{E}\left\|\mathbf{B}^\top\mathbf{B}\right\|_F^2 \le R^2\mathbb{E}\|\mathbf{B}\|_F^4 + 2R\mathbb{E}\|\mathbf{B}\|_F^4 = R(R+2)\mathbb{E}\|\mathbf{B}\|_F^4. \blacksquare$$

By using these lemmas and the exact same proof technique as in [29] one can find the bound for the variance (*e.g.* see section 5.1, bound on the variance of $f_{\mathrm{TT}(R)}$ part). ■

By employing Theorem 2, Theorem 5 in [29] and the hypercontractivity concentration inequality [30] we obtain the following theorem which leverages the bound on the variance to give a probabilistic bound on the RP's quality.

Theorem 5. *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 $\boldsymbol{\mathcal{X}} \in P$:*

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

Proof. The proof follows the one of Theorem 2 in [29] *mutatis mutandi*. □

4 Experiments

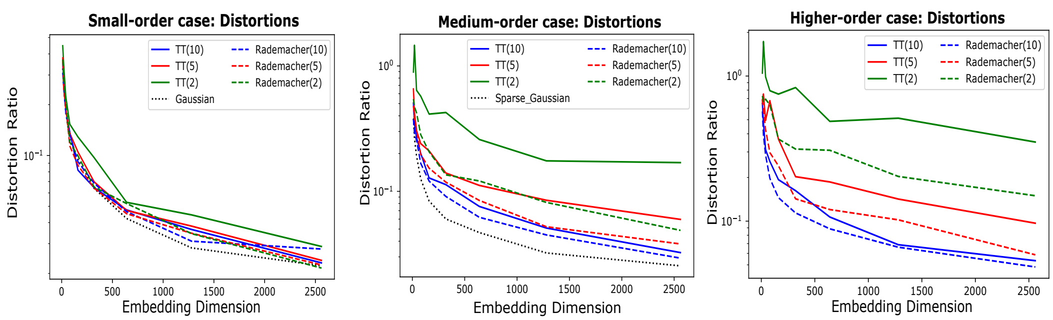


Figure 1: Comparison of the distortion ratio of tensorized Rademacher and tensorized Gaussian RPs and classical Gaussian RP for small-order (left), medium-order (center) and high-order (right) input tensors for different value of the rank parameter R.

We first compare the embedding performance of tensorized Rademacher and tensorized Gaussian RPs with classical Gaussian and very sparse [19] RPs on synthetic data for different size of input tensor and rank parameters. Second, to illustrate that the MPO RP used in [5] is not a well-suited dimension reduction map, we compare the Gaussian RP $f_{\mathrm{TT}(R)}$ proposed in [29] with the MPO RP defined in Section 3.1[1]. For both parts, the synthetic N-th order d dimensional tensor $\boldsymbol{\mathcal{X}}$ is generated in the TT format with the rank parameter equals to 10 with the entries of each core tensors drawn independently from the standard Gaussian distribution.

To compare tensorized Rademacher and Gaussian RPs, following [29] we consider three cases for different rank parameters: small-order ($d = 15, N = 3$), medium-order ($d = 3, N = 12$) and high-order ($d = 3, N = 25$). The embedding quality of each map is evaluated using the average distortion ratio $D(f, \boldsymbol{\mathcal{X}}) = \left| \frac{\|f(\boldsymbol{\mathcal{X}})\|^2}{\|\boldsymbol{\mathcal{X}}\|^2} - 1 \right|$ over 100 trials and is reported as a function of the projection size k in Figure 1. Note that due to memory requirements, the high order case cannot be handled with Gaussian or very sparse RPs. As we can see in the small-order case both tensorized maps perform competitively with classical Gaussian RP for all values of the rank parameter. In medium and high order cases, the quality of embedding of the tensorized Rademacher RP outperforms tensorized Gaussian RP for each value of the rank parameter. Moreover, the tensorized Rademacher RP gives us this speed up as there is no multiplication requirement in the calculations. This is shown in Figure 2 (right) where we report the time complexity of tensorized Rademacher RP vs tensorized Gaussian RP.

To validate the theoretical analysis in Proposition 1, we consider the medium-order case ($d = 3, N = 12$) and compare the Gaussian RP $f_{\mathrm{TT}(R)}$ with the MPO RP for different values of the rank parameter

[1]For these experiments we use *TT-Toolbox v2.2* [26].

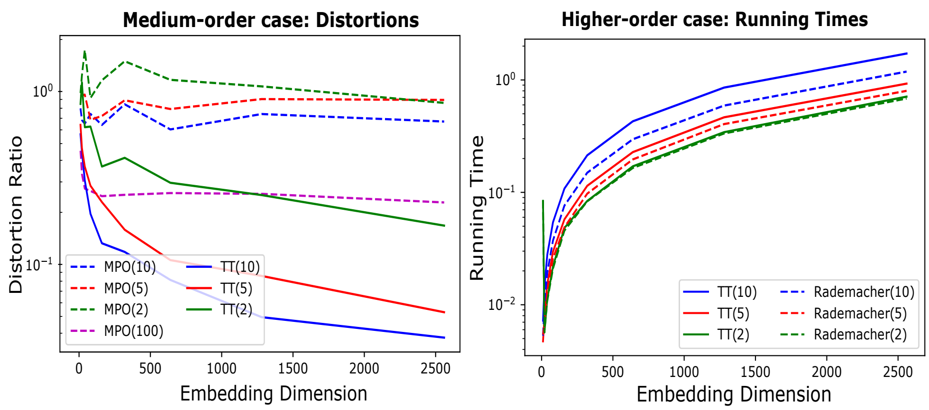


Figure 2: Comparison of distortion ratio of tensorized Gaussian RP and MPO RP for the medium-order case with different values for the rank parameter (left). Comparison of the running times between tensorized Rademacher and tensorized Gaussian RPs (right).

$R = 2, 5, 10$. These values correspond to roughly the same number of parameters that the two maps require. The quality of embedding via average distortion ratio over 100 trials is reported in Figure 2 where we see that the MPO RP performs really badly, which is predicted by our analysis. We also report the performance of the MPO RP with $R = 100$ to emphasize that the quality of the embedding does not reach acceptable levels even by greatly increasing the rank.

5 Conclusion

We presented an extension of the tensorized Gaussian embedding proposed in [29] for high-order tensors: Tensorized Rademacher random projection map. Our theoretical and empirical analysis show that the Gaussian tensorized RP in [29] can be replaced by the tensorized Rademacher RP while still benefiting from the JLT properties.We also showed, both in theory and practice, the RP in an MPO format is not a suitable dimension reduction map. Future research directions include leveraging and developing efficient sketching algorithms relying on tensorized RPs to find theoretical guarantees for randomized SVD and regression problems of very high-dimensional matrices given in the TT format.

Acknowledgment

This research was supported by the Canadian Institute for Advanced Research (CIFAR AI chair program) and the Natural Sciences and Engineering Research Council of Canada (Discovery program, RGPIN-2019-05949). BR was also supported by an IVADO Fellowship.

References

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

[2] Dimitris Achlioptas, Frank McSherry, and Bernhard Schölkopf. Sampling techniques for kernel methods. In *Advances in neural information processing systems*, pages 335–342, 2002.

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

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

[5] Kim Batselier, Wenjian Yu, Luca Daniel, and Ngai Wong. Computing low-rank approximations of large-scale matrices with the tensor network randomized svd. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1221–1244, 2018.

[6] 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.

[7] 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.

[8] Huaian Diao, Zhao Song, Wen Sun, and David Woodruff. Sketching for kronecker product regression and p-splines. In *International Conference on Artificial Intelligence and Statistics*, pages 1299–1308. PMLR, 2018.

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

[10] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

[11] Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang. Unsupervised generative modeling using matrix product states. *Physical Review X*, 8(3):031012, 2018.

[12] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[13] 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.

[14] 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.

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

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

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

[18] Peide Li. *Statistically Consistent Support Tensor Machine for Multi-Dimensional Data*. PhD thesis, Michigan State University, 2021.

[19] 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.

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

[21] Osman Asif Malik and Stephen Becker. Guarantees for the kronecker fast johnson–lindenstrauss transform using a coherence and sampling argument. *Linear Algebra and its Applications*, 602: 120–137, 2020.

[22] Jacob Miller, Guillaume Rabusseau, and John Terilla. Tensor networks for probabilistic sequence modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 3079–3087. PMLR, 2021.

[23] 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.

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

[25] Alexander Novikov, Mikhail Trofimov, and Ivan Oseledets. Exponential machines. *arXiv preprint arXiv:1605.03795*, 2016.

[26] Ivan Oseledets et al. Python implementation of the tt-toolbox. Available online, July 2015. URL `https://github.com/oseledets/ttpy`.

[27] Ivan V Oseledets. Approximation of $2^d \times 2^d$ matrices using tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2130–2145, 2010.

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

[29] Beheshteh Rakhshan and Guillaume Rabusseau. Tensorized random projections. In *International Conference on Artificial Intelligence and Statistics*, pages 3306–3316. PMLR, 2020.

[30] 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.

[31] 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.

[32] E Miles Stoudenmire. Learning relevant features of data with multi-scale tensor networks. *Quantum Science and Technology*, 3(3):034003, 2018.

[33] E Miles Stoudenmire and David J Schwab. Supervised learning with quantum-inspired tensor networks. *arXiv preprint arXiv:1605.05775*, 2016.

[34] 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.

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

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

[37] 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.