Guarantees for the Kronecker Fast
Johnson–Lindenstrauss Transform Using
a Coherence and Sampling Argument

Osman Asif Malik  
Department of Applied Mathematics 
University of Colorado Boulder  
osman.malik@colorado.edu

Stephen Becker  
Department of Applied Mathematics 
University of Colorado Boulder  
stephen.becker@colorado.edu

Abstract

In the recent paper [Jin, Kolda & Ward, arXiv:1909.04801], it is proved that the Kronecker fast Johnson–Lindenstrauss transform (KFJLT) is, in fact, a Johnson–Lindenstrauss transform, which had previously only been conjectured. In this paper, we provide an alternative proof of this, for when the KFJLT is applied to Kronecker vectors, using a coherence and sampling argument. Our proof yields a different bound on the embedding dimension, which can be combined with the bound in the paper by Jin et al. to get a better bound overall. As a stepping stone to proving our result, we also show that the KFJLT is a subspace embedding for matrices with columns that have Kronecker product structure. Lastly, we compare the KFJLT to four other sketch techniques in a numerical experiment.

1 Introduction

The Johnson–Lindenstrauss lemma, which was introduced by Johnson and Lindenstrauss [17], is the following fact.

**Theorem 1.1** (Johnson–Lindenstrauss lemma [10]). Let \( \varepsilon \in (0, 1) \) be a real number, let \( X \subseteq \mathbb{R}^I \) be a set of \( N \) points, and suppose \( J \geq C\varepsilon^{-2} \log N \), where \( C \) is an absolute constant. Then there exists a map \( f : \mathbb{R}^I \to \mathbb{R}^J \) such that for all \( x, y \in X \),

\[
(1 - \varepsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2.
\]

Any mapping \( f \) which has this property is called a Johnson–Lindenstrauss transform. Typically, such transforms are random maps, which motivates the following, more precise, definition.
**Definition 1.2** (Johnson–Lindenstrauss transform [27]). A probability distribution on a family of maps $F$, where each $f \in F$ maps $Y \subseteq \mathbb{R}^I$ to $\mathbb{R}^J$, is a *Johnson–Lindenstrauss transform* with parameters $\epsilon, \delta$, and $N$, or $\text{JLT}(\epsilon, \delta, N)$, on $Y$ if, for any subset $X \subseteq Y$ containing $N$ elements, the probability of drawing a map $f \in F$ which satisfies

$$
(\forall x, y \in X) \quad (1 - \epsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \epsilon)\|x - y\|_2^2
$$

is at least $1 - \delta$. Following common usage, we will refer to a random map as a JLT when the corresponding distribution satisfies this definition.

JLTs are usually constructed using simple random matrices, such as Gaussians with i.i.d. entries. They have many uses in applications, such as nearest neighbor searching [1], least squares regression [2, 13], sketching of data streams [27], and clustering [22].

When the vectors in the set $Y$ in Definition 1.2 have special structure, it is possible to construct a map $f$ that leverages this fact to speed up the computation of $f(x)$ when $x \in Y$. One class of vectors with such special structure are the Kronecker vectors $x = x^{(1)} \otimes x^{(2)} \otimes \cdots \otimes x^{(P)}$, where each $x^{(p)} \in \mathbb{R}^{I_p}$ and $\otimes$ denotes the Kronecker product. Vectors with Kronecker structure appear in various applications. When matricizing tensors in CP or Tucker format, the resulting matrices have columns which are Kronecker products. Computation with Kronecker vectors therefore feature in algorithms for computing these decompositions [18] and in related problems like tensor interpolative decomposition [7]. They also arise in areas such as higher dimensional numerical analysis [5, 6], tensor regression [11], and polynomial kernel approximation in machine learning [24]. The Kronecker fast Johnson–Lindenstrauss transform (KFJLT) is a map that can be applied very efficiently to Kronecker structured vectors. It was first proposed by Battaglino et al. [4] for solving the least squares problems that arise when computing the CP decomposition of tensors. Battaglino et al. [4] conjecture that the KFJLT is a JLT, but do not provide a proof. Recently, Jin et al. [16] provided a proof that the KFJLT indeed is a JLT.

In this paper, we provide an alternative proof of this fact for when the KFJLT is applied to Kronecker vectors, which is based on a coherence and sampling argument. As a stepping stone to proving our result, we also show that the KFJLT is an oblivious subspace embedding for matrices whose columns have Kronecker structure. Some ideas that we use in our proof were mentioned in [4]. Our guarantees are slightly different than those given in [16]: Ours have a worse dependence on the ambient dimensions $I_1, I_2, \ldots, I_P$ of the input vectors, but have a better dependence on the accuracy parameter $\epsilon$. The two bounds can be combined into one which yields a better bound overall. Another distinction between [16] and our paper is that the result in [16] shows that the KFJLT is a JLT on vectors with arbitrary structure, whereas our result is restricted to vectors with Kronecker structure.
2 Other Related Work

As mentioned in the introduction, a JLT can be constructed in many different ways. A popular choice is \( f(x) \triangleq \Omega x / \sqrt{J} \), where \( \Omega \in \mathbb{R}^{J \times I} \) has i.i.d. standard normal entries. More generally, the rows of \( \Omega \) can be chosen to be independent, mean zero, isotropic and sub-Gaussian random vectors in \( \mathbb{R}^I \) [26]. Ailon and Chazelle [1] proposed a fast JLT which leverages the Hadamard transform to achieve a transform that can be applied faster than a general dense matrix \( \Omega \).

A concept related to the JLT is subspace embedding.

**Definition 2.1** (Subspace embedding [27]). A \((1 \pm \epsilon)\) \(\ell_2\)-subspace embedding for the column space of a matrix \(X \in \mathbb{R}^{I \times R}\) is a matrix \(S \in \mathbb{R}^{J \times I}\) such that

\[
(\forall z \in \mathbb{R}^R) \quad (1 - \epsilon)\|Xz\|_2^2 \leq \|Sz\|_2^2 \leq (1 + \epsilon)\|Xz\|_2^2.
\]

A probability distribution on a family \(F\) of \(J \times I\) matrices is called an \((\epsilon, \delta)\) oblivious \(\ell_2\)-subspace embedding if the probability of drawing a matrix \(S \in F\) satisfying (1) is at least \(1 - \delta\). Following common usage, we will refer to a random matrix as an oblivious subspace embedding when the corresponding distribution satisfies this definition.

Thus, a subspace embedding distorts the squared length of a vector in the range of \(X\) by only a small amount. Methods for subspace embedding include leverage score sampling [20] and CountSketch [9]. Leverage score sampling is not an oblivious subspace embedding, since the sampling probabilities depend on \(X\). CountSketch, on the other hand, is an oblivious subspace embedding. Note that a subspace embedding is not necessarily a JLT. CountSketch, for example, is not a JLT [27]. For a more complete survey of work related to the JLT and subspace embedding, we refer the reader to the surveys in [21, 27].

For vectors with Kronecker structure, Sun et al. [25] propose the so called tensor random projection (TRP), whose transpose is a Khatri–Rao product of arbitrary random projection maps. They prove that TRP is a JLT in the special case when the TRP is constructed from two smaller random projections which have entries that are i.i.d. sub-Gaussians with zero mean and unit variance. The TRP idea is used in the earlier work [7] for tensor interpolative decomposition, but no guarantees are provided there.

Cheng et al. [8] propose an estimated leverage score sampling algorithm for \(\ell_2\)-regression when the design matrix is a Khatri–Rao product. They use this to speed up the alternating least squares algorithm for computing the tensor CP decomposition. A similar idea is proposed in Diao et al. [12] for \(\ell_2\)-regression when the design matrix is a Kronecker product.

The papers [23, 24, 3, 11] develop a method called TensorSketch, which is a variant of CountSketch that can be applied particularly efficiently to matrices whose columns have Kronecker structure. Avron et al. [3] show that TensorSketch is an oblivious subspace embedding, and Diao et al.
[11] provide guarantees for $\ell_2$-regression based on TensorSketch. However, just like CountSketch, TensorSketch is not a JLT.

3 Preliminaries

We use bold uppercase letters, e.g. $A$, to denote matrices; bold lowercase letters, e.g. $a$, to denote vectors; and regular lowercase letters, e.g. $a$, to denote scalars. Regular uppercase letters, e.g. $I, J, K$, are usually used to denote the size of vectors and matrices. This means that $I$ is a number and not the identity matrix. Subscripts are used to denote elements of matrices, and a colon denotes all elements in a row or column. For example, for a matrix $A$, $A_{ij}$ is the element on position $(i,j)$, $A_i$ is the $i$th row, and $A_{:j}$ is the $j$th column. Subscripts may be used to label different vectors. Superscripts in parentheses will be used for labeling both matrices and vectors. For example, $A^{(1)}$ and $A^{(2)}$ are two matrices. The norm $\|\cdot\|_2$ denotes the standard Euclidean norm for vectors, and the spectral norm for matrices. For matrices $A \in \mathbb{R}^{I \times J}$ and $B \in \mathbb{R}^{K \times L}$, their Kronecker product is denoted by $A \otimes B \in \mathbb{R}^{IK \times JL}$ and is defined as

$$A \otimes B \defeq \begin{bmatrix} A_{11}B & A_{12}B & \cdots & A_{1J}B \\ A_{21}B & A_{22}B & \cdots & A_{2J}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{IJ}B & A_{I2}B & \cdots & A_{IJ}B \end{bmatrix}.$$  

For matrices $A \in \mathbb{R}^{I \times K}$ and $B \in \mathbb{R}^{J \times K}$, their Khatri–Rao product is denoted by $A \odot B \in \mathbb{R}^{IJ \times K}$ and is defined as

$$A \odot B \defeq \begin{bmatrix} A_{11} \otimes B_{11} & A_{12} \otimes B_{12} & \cdots & A_{1K} \otimes B_{1K} \\ A_{21} \otimes B_{21} & A_{22} \otimes B_{22} & \cdots & A_{2K} \otimes B_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{I1} \otimes B_{I1} & A_{I2} \otimes B_{I2} & \cdots & A_{IK} \otimes B_{IK} \end{bmatrix}. $$

For an positive integer $n$, we use the notation $[n] \defeq \{1, 2, \ldots, n\}$. We let $\sigma_i(A)$ denote the $i$th singular value of the matrix $A$.

We now introduce the different tools we use to prove our results.

**Definition 3.1** (Randomized Hadamard transform [1]). Let $H \in \mathbb{R}^{I \times I}$ be the normalized Hadamard transform, and let $D \in \mathbb{R}^{I \times I}$ be a diagonal matrix with i.i.d. Rademacher random variables (i.e., equal to $+1$ or $-1$ with equal probability) on the diagonal. The $I \times I$ randomized Hadamard transform is defined as the random map $x \mapsto HDx$.

**Definition 3.2** (Leverage score, coherence [27]). Let $A \in \mathbb{R}^{I \times R}$ be a matrix, and let $\text{col}(A)$ be a matrix of size $I \times \text{rank}(A)$ whose columns form an orthonormal basis for range($A$). Then

$$\ell_i(A) \defeq \|\text{col}(A)_{i:}\|_2^2, \quad i \in [I],$$
is the $i$th leverage score of $A$. The coherence of $A$ is defined as

$$
\mu(A) \stackrel{\text{def}}{=} \max_{i \in [I]} \ell_i(A).
$$

The leverage scores, and consequently the coherence, do not depend on the particular basis chosen for the range of $A$ [27], so these quantities are well-defined. The coherence satisfies $\text{rank}(A)/I \leq \mu(A) \leq 1$.

**Definition 3.3** (Leverage score sampling [27]). Let $A \in \mathbb{R}^{I \times R}$ and $p_i \stackrel{\text{def}}{=} \ell_i(A)/\text{rank}(A)$ for all $i \in [I]$. Then $p \stackrel{\text{def}}{=} [p_1, p_2, \ldots, p_I]$ is a probability distribution on $[I]$. Let $q \stackrel{\text{def}}{=} [q_1, q_2, \ldots, q_I]$ be another probability distribution on $[I]$, and suppose that for some $\beta \in (0, 1]$ it satisfies $q_i \geq \beta p_i$ for all $i \in [I]$. Let $v \in [I]^J$ be a random vector with independent elements satisfying $\mathbb{P}(v_j = i) = q_i$ for all $(i, j) \in [I] \times [J]$. Let $\Omega \in \mathbb{R}^{J \times I}$ and $R \in \mathbb{R}^{J \times J}$ be a random sampling matrix and a diagonal rescaling matrix, respectively, defined as

$$
\Omega_j = e_{v_j}^\top \quad \text{and} \quad R_{jj} = \frac{1}{\sqrt{jq_{v_j}}}
$$

for each $j \in [J]$, where $e_i$ is the $i$th column of the $I \times I$ identity matrix. The leverage score sampling matrix $S_q \in \mathbb{R}^{J \times I}$ is then defined as $S_q \stackrel{\text{def}}{=} R\Omega$, where the subscript indicates that the sampling is done according to the distribution $q$.

**Definition 3.4** (Kronecker fast Johnson–Lindenstrauss transform [16]). For each $p \in [P]$, let $H^{(p)}D^{(p)}$ be independent randomized Hadamard transforms of size $I_p \times I_p$. The **Kronecker fast Johnson–Lindenstrauss transform** (KFJLT) of a vector $x = x^{(1)} \otimes x^{(2)} \otimes \cdots \otimes x^{(P)}$, with $x^{(p)} \in \mathbb{R}^{I_p}$, is defined as

$$
S_q \left( \bigotimes_{p=1}^P H^{(p)}D^{(p)} \right) x = S_q \left( \bigotimes_{p=1}^P H^{(p)}D^{(p)}x^{(p)} \right),
$$

(2)

where $S_q \in \mathbb{R}^{J \times \hat{I}}$ is a sampling matrix as in Definition 3.3 with $q$ equal to the uniform distribution on $[\hat{I}]$, where $\hat{I} \stackrel{\text{def}}{=} I_1I_2 \cdots I_P$. The equality in (2) follows from a basic property of the Kronecker product; see e.g. Lemma 4.2.10 in [15].

A benefit of the KFJLT is that the Kronecker structured vector does not have to be explicitly computed—it is sufficient to store the smaller vectors $x^{(1)}, x^{(2)}, \ldots, x^{(P)}$. Another benefit is that each randomized Hadamard transform $H^{(p)}D^{(p)} \in \mathbb{R}^{I_p \times I_p}$ only costs $O(I_p \log I_p)$ to apply to $x^{(p)}$.

Lemma 3.5 below is a variant of Lemma 3 in [13] but with an arbitrary probability of success. The proof is identical to that for Lemma 3 in [13]—which in turn follows similar reasoning as in the proof of Lemma 1 in [1]—but using an arbitrary failure probability $\eta$ instead of $1/20$, combined with the definition of leverage score and coherence.

---

1. Jin et al. [16] use the discrete Fourier transform instead of the Hadamard transform in their definition.
Lemma 3.5. Let $A \in \mathbb{R}^{I \times R}$ be a matrix and let $HD$ be the $I \times I$ randomized Hadamard transform. Then, with probability at least $1 - \eta$, the following holds:

$$\mu(HDA) \leq \frac{2R \ln(2IR/\eta)}{I}.$$  

Lemma 3.6 below is a restated version of Theorem 3.3 in [8]. A similar statement is also made in Lemma 4 in [4].

Lemma 3.6. For each $p \in [P]$, let $A^{(p)} \in \mathbb{R}^{I_p \times R}$. Then

$$\mu\left(\bigotimes_{p=1}^{P} A^{(p)}\right) \leq \prod_{p=1}^{P} \mu(A^{(p)}).$$

Lemma 3.7 below is a slight restatement of Theorem 2.11 in [27], with a careful choice of the constant parameter.

Lemma 3.7. Let $A \in \mathbb{R}^{I \times R}$ and assume $\varepsilon \in (0, 1)$. Suppose

$$J > \frac{8R \ln(2R/\eta)}{3 \beta \varepsilon^2}$$

and that $S_q \in \mathbb{R}^{J \times 1}$ is a leverage score sampling matrix as in Definition 3.3, where the $\beta$ in that definition is the same as the $\beta$ in (3). Then, with probability at least $1 - \eta$, the following holds:

$$(\forall i \in \text{rank}(A)) \quad 1 - \varepsilon \leq \sigma_i^2(S_q \text{col}(A)) \leq 1 + \varepsilon.$$  

4 Main Results

We consider the set

$$\mathcal{Y} = \{x \in \mathbb{R}^I : x = x^{(1)} \otimes x^{(2)} \otimes \cdots \otimes x^{(P)}, \text{ with } x^{(p)} \in \mathbb{R}^{I_p} \text{ for each } p \in [P]\}$$

of Kronecker vectors. Theorem 4.1 shows that the KFJLT is an $(\varepsilon, \delta)$ oblivious $\ell_2$-subspace embedding for matrices whose columns have Kronecker product structure when the embedding dimension $J$ is sufficiently large.

Theorem 4.1. Let $X = [x_1, x_2, \ldots, x_R] \in \mathbb{R}^{I \times R}$ be a matrix with each column $x_r = \bigotimes_{p=1}^{P} x^{(p)}_r \in \mathcal{Y}$. For each $p \in [P]$, let $H^{(p)}D^{(p)}$ be independent randomized Hadamard transforms of size $I_p \times I_p$, and

---

2The statement in [27] has a constant 144 instead of $8/3$. However, we found that $8/3$ is sufficient under the assumption that $\varepsilon \in (0, 1)$. The proof given in [27] otherwise remains the same.
define
\[ \Phi \overset{\text{def}}{=} \bigotimes_{p=1}^{P} H^{(p)} D^{(p)}. \]

Let \( S_q \in \mathbb{R}^{J \times \tilde{I}} \) be a sampling matrix as in Definition 3.3 with \( q \) equal to the uniform distribution, and assume \( \varepsilon \in (0, 1) \). If

\[ J > \frac{8}{3} \cdot 2^P R^{P+1} \varepsilon^{-2} \ln \left( \frac{2R(P + 1)}{\delta} \right) \prod_{p=1}^{P} \ln \left( \frac{2I_p R(P + 1)}{\delta} \right), \]

then the following holds with probability at least \( 1 - \delta \):

\[ (\forall z \in \mathbb{R}^R) \quad (1 - \varepsilon) \|Xz\|_2^2 \leq \|S_q \Phi Xz\|_2^2 \leq (1 + \varepsilon) \|Xz\|_2^2. \]

**Proof.** If all columns of \( X \) are the zero vector, the claim is trivially true. So we now assume that at least one column of \( X \) is nonzero. Note that \( X = \bigotimes_{p=1}^{P} X^{(p)} \), where each \( X^{(p)} \) def \(= [x^{(p)}_1, x^{(p)}_2, \ldots, x^{(p)}_{R_p}] \).

By Lemma 3.5, for a fixed \( p \in [P] \), the following holds with probability at least \( 1 - \eta \):

\[ \mu(H^{(p)} D^{(p)} X^{(p)}) \leq \frac{2R \ln(2I_p R/\eta)}{I_p}. \]

Hence, taking a union bound, the following holds with probability at least \( 1 - P\eta \):

\[ (\forall p \in [P]) \quad \mu(H^{(p)} D^{(p)} X^{(p)}) \leq \frac{2R \ln(2I_p R/\eta)}{I_p}. \]

Now applying Lemma 3.6, we have that the following holds with probability at least \( 1 - P\eta \):

\[ \mu(\Phi X) = \mu(\bigotimes_{p=1}^{P} H^{(p)} D^{(p)} X^{(p)}) \leq \prod_{p=1}^{P} \mu(H^{(p)} D^{(p)} X^{(p)}) \leq \frac{1}{I} \prod_{p=1}^{P} 2R \ln(2I_p R/\eta). \]

(5)

For \( i \in [\tilde{I}] \), let \( p_i = \ell_i(\Phi X)/\text{rank}(\Phi X) \). Since \( \Phi \) is a Kronecker product of orthogonal matrices, \( \Phi \) is also orthogonal [19], and since \( X \) is nonzero, it follows that \( \text{rank}(\Phi X) \geq 1 \), so \( p_i \) is well defined. Instead of sampling according to the unknown distribution \([p_1, p_2, \ldots, p_I]\), we sample according to the uniform distribution \( q = [q_1, q_2, \ldots, q_I] \). To get guarantees, we want to apply Lemma 3.7. To do this, we first need to find some \( \beta \in (0, 1] \) such that

\[ (\forall i \in [\tilde{I}]) \quad q_i = \frac{1}{I} \geq \beta p_i. \]
From (5), the following holds with probability at least $1 - P\eta$:

$$p_i = \frac{\ell_i(\Phi X)}{\operatorname{rank}(\Phi X)} \leq \mu(\Phi X) \leq \frac{1}{\tilde{I}} \prod_{p=1}^{P} 2R \ln(2I_p R/\eta).$$

Hence, choosing $\beta$ such that

$$\beta^{-1} = \prod_{p=1}^{P} 2R \ln(2I_p R/\eta)$$

ensures that $q_i = 1/\tilde{I} \geq \beta p_i$ for all $i \in [\tilde{I}]$ with probability at least $1 - P\eta$. Let $\alpha \overset{\text{def}}{=} \operatorname{rank}(\Phi X) = \operatorname{rank}(X)$, and let $U \Sigma V^T = X$ be the SVD of $X$ with $U \in \mathbb{R}^{\tilde{I} \times \alpha}$, $\Sigma \in \mathbb{R}^{\alpha \times \alpha}$ and $V \in \mathbb{R}^{R \times \alpha}$. Note that the columns of $\Phi U$ form an orthonormal basis for range($\Phi X$). Hence, we can choose $\operatorname{col}(\Phi X) = \Phi U$. Using Lemma 3.7, with $A = \Phi X \in \mathbb{R}^{\tilde{I} \times R}$, it follows that if

$$J > \frac{8}{3} \cdot 2^P R^{P+1} \varepsilon^{-2} \ln(2R/\eta) \prod_{p=1}^{P} \ln(2I_p R/\eta),$$

then the following holds with probability at least $1 - (P + 1)\eta$:

$$(\forall i \in [\alpha]) \quad 1 - \varepsilon \leq \sigma_i^2(S_q \Phi U) \leq 1 + \varepsilon.$$ 

By the minimax characterization of singular values (see e.g. Theorem 8.6.1 in [14]), it follows that

$$(\forall w \in \mathbb{R}^\alpha) \quad (1 - \varepsilon) \|w\|_2^2 \leq \|S_q \Phi U w\|_2^2 \leq (1 + \varepsilon) \|w\|_2^2$$

holds with probability at least $1 - (P + 1)\eta$. In particular, for any $z \in \mathbb{R}^R$, this is true for $w = \Sigma V^T z \in \mathbb{R}^\alpha$. Consequently,

$$(\forall z \in \mathbb{R}^R) \quad (1 - \varepsilon) \|\Sigma V^T z\|_2^2 \leq \|S_q \Phi U \Sigma V^T z\|_2^2 \leq (1 + \varepsilon) \|\Sigma V^T z\|_2^2,$$

or equivalently,

$$(\forall z \in \mathbb{R}^R) \quad (1 - \varepsilon) \|X z\|_2^2 \leq \|S_q \Phi X z\|_2^2 \leq (1 + \varepsilon) \|X z\|_2^2,$$

holds with probability at least $1 - (P + 1)\eta = 1 - \delta$, where $\delta \overset{\text{def}}{=} (P + 1)\eta$. Replacing $\eta = \delta/(P + 1)$ in (6) gives (4).

The following theorem is our main result. It shows that the KFJLT is a JLT($\varepsilon, \delta, N$) on $\mathcal{X}$ when the embedding dimension $J$ is sufficiently large.

**Theorem 4.2.** Let $X \subseteq \mathcal{X}$ consist of $N$ distinct vectors with Kronecker structure. Let $\Phi$ be defined as in Theorem 4.1. Let $S_q \in \mathbb{R}^{J \times I}$ be a sampling matrix as in Definition 3.3 with $q$ equal to the
uniform distribution, and assume \( \varepsilon \in (0, 1) \). If

\[
J > \frac{16}{3} \cdot 4^P \varepsilon^{-2} \ln \left( \frac{4N^2(P+1)}{\delta} \right) \prod_{p=1}^{P} \ln \left( \frac{4I_pN^2(P+1)}{\delta} \right),
\]

then the following holds with probability at least 1 - \( \delta \):

\[
(\forall x, y \in \mathcal{X}) \quad (1 - \varepsilon)\|x - y\|_2^2 \leq \|S_q\Phi x - S_q\Phi y\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2.
\]

Proof. Fix \( x, y \in \mathcal{X} \) and set \( X \overset{\text{def}}{=} [x, y] \). From Theorem 4.1, we know that if

\[
J > \frac{8}{3} \cdot 2^P 2^{P+1} \varepsilon^{-2} \ln \left( \frac{4(P+1)}{\eta} \right) \prod_{p=1}^{P} \ln \left( \frac{4I_p(P+1)}{\eta} \right),
\]

then the following holds with probability at least 1 - \( \eta \):

\[
(\forall z \in \mathbb{R}^2) \quad (1 - \varepsilon)\|Xz\|_2^2 \leq \|S_q\Phi Xz\|_2^2 \leq (1 + \varepsilon)\|Xz\|_2^2.
\]

In particular, setting \( z = [1, -1]^\top \), we have that, with probability at least 1 - \( \eta \),

\[
(1 - \varepsilon)\|x - y\|_2^2 \leq \|S_q\Phi x - S_q\Phi y\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2.
\]

Taking a union bound over all distinct \( N^2 - N \) pairs \( x, y \in \mathcal{X} \), we have that (8) holds with probability at least 1 - \( N^2 \eta = 1 - \delta \), where \( \delta \overset{\text{def}}{=} N^2 \eta \). Replacing \( \eta = \delta/N^2 \) in (9) gives (7).

Assuming \( N > \max\{P,4\} \), the bound in (7) can be simplified to

\[
J > C_1 \varepsilon^{-2} C_2^P \log \left( \frac{N}{\delta} \right) \prod_{p=1}^{P} \log \left( \frac{I_pN}{\delta} \right),
\]

where \( C_1 \) and \( C_2 \) are absolute constants. For comparison, and expressed in the same notation as in this paper, the bound on \( J \) in Theorem 2.1 in [16] needed to guarantee that the KFJLT is a JLT(\( \varepsilon, \delta, N \)) on \( \mathbb{R}^I \) is of the form

\[
J > C \varepsilon^{-2} \log^{2P-1} \left( \frac{N}{\delta} \right) \log^4 \left( \varepsilon^{-1} \log \left( \frac{N}{\delta} \right) \right) \log \left( \prod_{p=1}^{P} I_p \right),
\]

where \( C \) is an absolute constant. The bound in (11) has a nicer dependence on the dimension sizes
$I_1, I_2, \ldots, I_P$ than the bound in (10). Indeed, the term
\[
\log \left( \prod_{p=1}^{P} I_p \right) = \sum_{p=1}^{P} \log(I_p)
\]
in (11) is a sum of logs of $I_1, I_2, \ldots, I_P$, whereas the term
\[
\prod_{p=1}^{P} \log \left( \frac{I_p N}{\delta} \right)
\]
in (10) is a product of logs of $I_1 N/\delta, I_2 N/\delta, \ldots, I_P N/\delta$. On the other hand, the bound in (10) has a nicer dependence on $\varepsilon$ than (11) does. Indeed, (10) contains the term $(\varepsilon^{-2})$ whereas (11) contains the term $(\varepsilon^{-2} \log^4(\varepsilon^{-1}))$. These two bounds can therefore be combined to yield a better bound on the size of $J$ required to ensure that the KFJLT is a JLT($\varepsilon, \delta, N$) on $Y$.

5 Numerical Experiments

In this section, we present the results from an experiment which compares five different sketches when applied to random Kronecker vectors with three different random distributions\(^3\). The five methods we compare are the following.

- **Gaussian** sketch uses an unstructured $J \times \bar{I}$ matrix with i.i.d. standard normal entries which are scaled by $1/\sqrt{J}$. This approach is not scalable, but interesting to use as a baseline in this experiment.

- **KFJLT** is the sketch discussed in this paper, and which is defined in Definition 3.4, with the only difference that the uniform sampling of rows is done without replacement.

- **TRP** is the method proposed in [25]. As sub-matrices, we use matrices with i.i.d. standard normal entries of size $I_p \times J$ and rescale appropriately.

- **TensorSketch** is the method developed in [23, 24, 3, 11].

- **Sampling** is the method proposed by [8]. It computes an estimate of the leverage scores for each row of the matrix to be sampled and uses these to compute a distribution which is used for sampling.

All of these methods, except the first, are specifically designed for sketching vectors with Kronecker structure. As input, we use random Kronecker vectors $\mathbf{x} = \bigotimes_{p=1}^{3} \mathbf{x}^{(p)}$, where each $\mathbf{x}^{(p)} \in \mathbb{R}^{16}$ has one of the following distributions.

\(^3\)The code used for this experiment is available at [https://github.com/OsmanMalik/kronecker-sketching](https://github.com/OsmanMalik/kronecker-sketching).
• Each \( x^{(p)} \) has i.i.d. standard normal entries.

• Each \( x^{(p)} \) is sparse with three nonzero elements, which are independent and normally distributed with mean zero and standard deviation 100. The positions of the three nonzero elements are drawn uniformly at random without replacement.

• Each \( x^{(p)} \) contains a single nonzero entry, which is chosen uniformly at random. This nonzero entry is equal to 100.

Sparse Kronecker vectors are interesting in many data science applications, and arise in decomposition of sparse tensors, for example.

In the experiment, we draw two random vectors \( x, y \in \mathbb{R}^{4096} \), each of which is a Kronecker product of three smaller vectors of length 16, drawn according to one of the three distributions above. For each of the five sketches \( f \) and for some embedding dimension \( J \), we then compute how well they preserve the distance between \( x \) and \( y \) by computing the quantity

\[
\frac{\| f(x) - f(y) \|_2}{\| x - y \|_2} - 1.
\]

(12)

For each of the three distributions and each embedding dimension

\[ J \in \{100, 200, \ldots, 1000\} \]

we repeat this 1000 times and compute the mean, standard deviation and maximum of (12) over those 1000 trials. Figures 1–3 present the results for each of the three distributions.

Figure 1: Mean, standard deviation, and maximum of the quantity in (12) over 1000 trials when the test vectors are Kronecker products of vectors with i.i.d. standard normal entries.

No one method produces the best results for all three distributions. The leverage score sampling approach does very well on dense vectors, even outperforming the Gaussian sketch, but does less well on sparser inputs. Although TensorSketch has an impressive mean performance on the two
Figure 2: Mean, standard deviation, and maximum of the quantity in (12) over 1000 trials when the test vectors are Kronecker products of vectors with three nonzero elements which are independent and normally distributed with mean zero and standard deviation 100.

Figure 3: Mean, standard deviation, and maximum of the quantity in (12) over 1000 trials when the test vectors are Kronecker products of vectors containing a single nonzero entry equal to 100.

sparser inputs, it sometimes produces high distortion rates on those inputs. On the sparser inputs, the KFJLT seems to strike the best balance between mean and worst case performance. TRP does poorly for all three distribution types.

6 Conclusion

We have presented a coherence and sampling argument for showing that the KFJLT is a Johnson–Lindenstrauss transform on vectors with Kronecker structure. Since our bound on the embedding dimension is different from the one in the recent paper by Jin et al. [16], it can be combined with the bound from that paper to yield a better bound overall. As a stepping stone to proving our result, we also showed that the KFJLT is a subspace embedding for matrices whose columns are Kronecker products.
We provided results from a numerical experiment which compares five different sketches, four of which are designed to be particularly efficient for sketching of Kronecker structured vectors. The experiments were done on Kronecker vectors with three different random distributions. No single method outperformed all others for all three vector distributions. This indicates that different sketch techniques may work best in different applications. We believe that there is a need for a more comprehensive comparison of sketches for structured data to help practitioners choose the best sketch for their particular needs.

Acknowledgments

We would like to thank Tammy Kolda, Ruhui Jin and Rachel Ward for providing feedback on an early version of this paper.

This material is based upon work supported by the National Science Foundation under Grant No. ECCS-1810314.

References

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

[2] Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging LAPACK’s least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

[3] Haim Avron, Huy L. Nguyen, and David P. Woodruff. Subspace Embeddings for the Polynomial Kernel. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2258–2266, Cambridge, MA, USA, 2014. MIT Press.

[4] Casey Battaglino, Grey Ballard, and Tamara G. Kolda. A practical randomized CP tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.

[5] Gregory Beylkin and Martin J. Mohlenkamp. Numerical operator calculus in higher dimensions. *Proceedings of the National Academy of Sciences*, 99(16):10246–10251, 2002.

[6] Gregory Beylkin and Martin J. Mohlenkamp. Algorithms for Numerical Analysis in High Dimensions. *SIAM Journal on Scientific Computing*, 26(6):2133–2159, July 2006.

[7] David J. Biagioni, Daniel Beylkin, and Gregory Beylkin. Randomized interpolative decomposition of separated representations. *Journal of Computational Physics*, 281(C):116–134, January 2015. ISSN 0021-9991. doi: 10.1016/j.jcp.2014.10.009.
[8] Dehua Cheng, Richard Peng, Yan Liu, and Ioakeim Perros. SPALS: Fast alternating least squares via implicit leverage scores sampling. In Advances In Neural Information Processing Systems, pages 721–729, 2016.

[9] Kenneth L. Clarkson and David P. Woodruff. Low-Rank Approximation and Regression in Input Sparsity Time. Journal of the ACM, 63(6):54:1–54:45, February 2017. ISSN 0004-5411. doi: 10.1145/3019134.

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

[11] Huaian Diao, Zhao Song, Wen Sun, and David Woodruff. Sketching for Kronecker Product Regression and P-splines. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, pages 1299–1308, 2018.

[12] Huaian Diao, Rajesh Jayaram, Zhao Song, Wen Sun, and David P. Woodruff. Optimal Sketching for Kronecker Product Regression and Low Rank Approximation. arXiv preprint arXiv:1909.13384, 2019.

[13] Petros Drineas, Michael W. Mahoney, Shan Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. Numerische mathematik, 117(2):219–249, 2011.

[14] Gene H. Golub and Charles F. Van Loan. Matrix Computations. Johns Hopkins University Press, Baltimore, 4th edition, 2013. ISBN 978-1-4214-0794-4.

[15] Roger A. Horn and Charles R. Johnson. Topics in Matrix Analysis. Cambridge University Press, 1994.

[16] Ruhui Jin, Tamara G. Kolda, and Rachel Ward. Faster Johnson-Lindenstrauss Transforms via Kronecker Products. arXiv preprint arXiv:1909.04801, 2019.

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

[18] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. SIAM Review, 51(3):455–500, August 2009. ISSN 0036-1445. doi: 10.1137/07070111X.

[19] Charles F. Van Loan. The ubiquitous Kronecker product. Journal of Computational and Applied Mathematics, 123(1):85–100, November 2000. ISSN 0377-0427. doi: 10.1016/S0377-0427(00)00393-9.

[20] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative Bernstein bound. arXiv preprint arXiv:1008.0587, 2010.
[21] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[22] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of Johnson–Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038. ACM, 2019.

[23] Rasmus Pagh. Compressed Matrix Multiplication. *ACM Transactions on Computation Theory*, 5(3):9:1–9:17, August 2013. ISSN 1942-3454. doi: 10.1145/2493252.2493254.

[24] 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*, KDD ’13, pages 239–247, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487591.

[25] 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.

[26] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[27] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(12):1–157, 2014.