SAND2019-11639C

Faster Johnson-Lindenstrauss Transforms via Kronecker Products *

Ruhui Jin†, Tamara G. Kolda‡, and Rachel Ward†

Abstract. The Kronecker product is an important matrix operation with a wide range of applications in supporting fast linear transforms, including signal processing, graph theory, quantum computing and deep learning. In this work, we introduce a generalization of the fast Johnson-Lindenstrauss projection for embedding vectors with Kronecker product structure, the *Kronecker fast Johnson-Lindenstrauss transform* (KFJLT). The KFJLT drastically reduces the embedding cost to an exponential factor of the standard fast Johnson-Lindenstrauss transform (FJLT)'s cost when applied to vectors with Kronecker structure, by avoiding explicitly forming the full Kronecker products. We prove that this computational gain comes with only a small price in embedding power: given $N = \prod_{k=1}^{d} n_k$, consider a finite set of p points in a tensor product of d constituent Euclidean spaces $\bigotimes_{k=d}^{1} \mathbb{R}^{n_k} \subset \mathbb{R}^N$. With high probability, a random KFJLT matrix of dimension $N \times m$ embeds the set of points up to multiplicative distortion $(1 \pm \varepsilon)$ provided by $m \gtrsim \varepsilon^{-2} \cdot \log^{2d-1}(p) \cdot \log N$. We conclude by describing a direct application of the KFJLT to the efficient solution of large-scale Kronecker-structured least squares problems for fitting the CP tensor decomposition.

Key words. Johnson-Lindenstrauss embedding, fast Johnson-Lindenstrauss transform (FJLT), Kronecker structure, concentration inequality, restricted isometry property.

1. Introduction.

Dimensionality reduction is commonly used in data analysis to project high-dimensional data onto a lower-dimensional space while preserving as much information as possible. The powerful *Johnson-Lindenstrauss lemma* proves the existence of a class of linear maps which provide low-distortion embeddings of an arbitrary number of points from high-dimensional Euclidean space into a exponentially lower dimensional space [25, 15].

A (distributional) Johnson-Lindenstrauss transform (JLT) is a random linear map which provides such an embedding with high probability, and a *fast* JL transform (FJLT) exploiting fast matrix-vector multiplies of the FFT significantly reduces the complexity of the embedding with only a minor increase in the embedding dimension [1, 2, 3]. We consider the dimensionality reduction problem for high-dimensional subspaces with *structure*, specifically, subspaces corresponding to a tensor product of lower-dimensional Euclidean spaces. In this case, we can dramatically reduce the embedding complexity with only a small increase in the embedding dimension (see Figure 1).

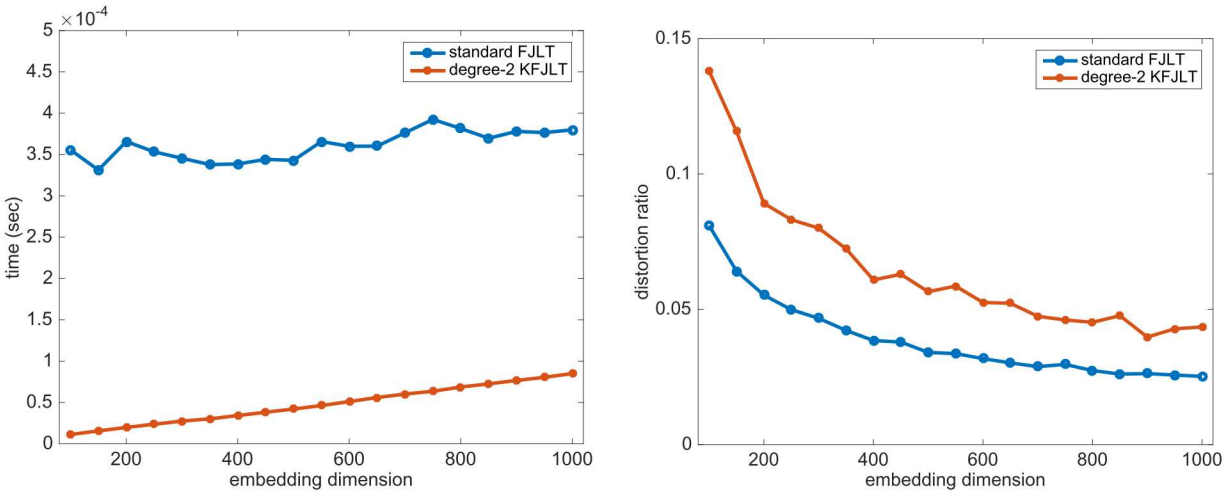
1.1. Review of JLT and FJLT.

We briefly review JLT and FJLT. Suppose we have a set $\mathcal{E} \subset \mathbb{R}^N$ of p points. A JLT is a (random) linear map $\mathbf{\Phi}$ from $\mathbb{R}^N$ down to $\mathbb{R}^m$ with ideally $m_{\text{opt}} = O(\varepsilon^{-2} \cdot \log p)$ [25, 28] such that with high probability with respect to the draw of $\mathbf{\Phi}$,

*R. Jin and R. Ward are supported by AFOSR MURI Award N00014-17-S-F006. T. G. Kolda is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program. Sandia National Laboratories is a multimission laboratory manage and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

†Department of Mathematics, University of Texas, Austin, TX (rhjin@math.utexas.edu,rward@math.utexas.edu)

‡Sandia National Laboratories, Livermore, CA (tgkolda@sandia.gov)



(a) Embedding time (b) Embedding distortion

Figure 1. *Comparison of the embedding time and distortion between a standard FJLT and a Kronecker FJLT on one vector with Kronecker structure* $\mathbb{R}^{125} \otimes \mathbb{R}^{125}$ *using* MATLAB R2015a *fft() and Tensor Toolbox* v3.1 *[5] on a standard MacBook Pro* 2016 *with* 16 *GB of memory. Each dot in the vertical direction of represents the average embedding time and distortion for a given embedding dimension on the same bulk of* 1000 *appropriately structured Kronecker vectors. Each component vector consist of normally distributed elements.*

the transformed points have at most $(1 \pm \varepsilon)$ multiplicative distortion, i.e.,

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2 \quad \text{for all} \quad \mathbf{x} \in \mathcal{E}. \tag{1.1}$$

The Gaussian testing matrix

$$\mathbf{\Phi} \in \mathbb{R}^{m \times N} \quad \text{with} \quad m = O(\varepsilon^{-2} \cdot \log p) \quad \text{and with i.i.d.} \quad \phi_{ij} \sim \mathcal{N}(0, 1/m) \tag{1.2}$$

is a particular JLT which achieves optimal distortion power.

Although this is a powerful result, the cost of this transformation for each point is $O(mN)$. To reduce the cost, *fast* JLTs (FJLTs) employ fast matrix-vector multiplication [1, 2, 3]. An example FJLT is of the form

$$\begin{aligned} \mathbf{\Phi} = \mathbf{S}\mathcal{F}_N\mathbf{D}_N \in \mathbb{C}^{m\times N} \quad \text{with} \quad m_{\mathrm{f}} = O\Big(\varepsilon^{-2} \cdot \log p \cdot \log^4(\log p) \cdot \log N\Big) \text{ [27, 23]}, \\ \mathbf{S} \in \mathbb{R}^{m\times N} = m \text{ random rows of the } N \times N \text{ identity matrix}, \\ \mathcal{F}_N \in \mathbb{C}^{N\times N} = \text{fast Fourier transform of dimension } N, \text{ and} \\ \mathbf{D}_N \in \mathbb{R}^{N\times N} = \text{diagonal matrix with } N \text{ random } \pm 1 \text{ entries}. \end{aligned} \tag{1.3}$$

Note that the embedding dimension, m_{f}, in the FJLT is increased by the small factor of $\log^2(\log p) \cdot \log N$ as compared to the optimal JLT. However, the per-point transformation cost is reduced from $O(mN)$ to $O(N \log N + m)$.

1.2. Our contribution: Kronecker FJLT. In this work, we consider the following scenario. We assume that our set $\mathcal{E}$ of p points comes from a tensor product space, i.e.,

$$\mathcal{E} \subset \bigotimes_{k=d}^{1} \mathbb{R}^{n_k} \subset \mathbb{R}^N \quad \text{where} \quad N = \prod_{k=1}^{d} n_k.$$

Each vector $\mathbf{x} \in \mathcal{E}$ is a Kronecker product which means that each entry of $\mathbf{x}$ is the product of entries of the constituent vectors:

$$\mathbf{x} = \bigotimes_{k=d}^{1} \mathbf{x}_k \in \mathbb{R}^N \text{ where } \mathbf{x}_k \in \mathbb{R}^{n_k}, \quad \text{i.e.,} \quad x(i) = \prod_{k=d}^{1} x_k(i_k) \text{ where } i = 1 + \sum_{k=1}^{d} (i_k - 1) \prod_{\ell=1}^{k-1} n_\ell.$$

The Kronecker product is an important matrix operation with a wide range of applications in supporting fast linear transforms [37, 34], signal processing [18, 17], graph theory [29], quantum computing [19], deep learning [30] and so on.

For a set of *Kronecker* structured vectors, we propose a *Kronecker* FJLT (KFJLT) of the form

(1.4)

$$\mathbf{\Phi} = \mathbf{S} \bigotimes_{k=d}^{1} (\mathcal{F}_{n_k} \mathbf{D}_{n_k}) \in \mathbb{C}^{m \times N} \quad \text{with} \quad m_{\text{kron}} = O\Big(\varepsilon^{-2} \cdot \log^{2d-1}(p) \cdot \log^4(\log p) \cdot \log N\Big),$$

$$\begin{aligned} \mathbf{S} \in \mathbb{R}^{m \times N} &= m \text{ random rows of the } N \times N \text{ identity matrix,} \\ \mathcal{F}_{n_k} \in \mathbb{C}^{n_k \times n_k} &= \text{fast Fourier transform of dimension } n_k, \text{ and} \\ \mathbf{D}_{n_k} \in \mathbb{R}^{n_k \times n_k} &= \text{diagonal matrix with } n_k \text{ random } \pm 1 \text{ entries.} \end{aligned}$$

The $\mathbf{S}$ matrix is unchanged, but $\mathcal{F}_N \mathbf{D}_N$ has been replaced by a Kronecker product. We call d the *degree* of the KFJLT.

For Kronecker-structured vectors, the KFJLT reduces the transformation cost of each point to $O(\sum_{k=1}^{d} n_k \log n_k + m)$. As compared to the FJLT, the necessary embedding dimension m has increased by only a factor of $\log^{2d-2}(p)$ (when $d = 1$, the KFJLT reduces to the standard FJLT). This idea was proposed in the context of matrix sketching for the least squares problems in fitting the CANDECOMP/PARAFAC (CP) tensor decomposition [7]; however, there was no proof such a transform was a JLT. In this work, we prove that this is a JLT and that the embedding dimension is only slightly worse than in the FJLT case.

1.3. Related work. Sun et al. [36] proposed a related tensor-product embedding construction called the tensor random projection (TRP). The TRP is a low-memory framework for random maps formed by a row-wise Kronecker product of common embedding matrices; for example, Gaussian testing matrices and sparse random projections. The authors provide theoretical analysis for the case of the component random maps being two Gaussian matrices. Our theoretical embedding results are favorable to theirs in several key aspects: our embedding bound applies to fast JLTs which support fast matrix multiplications, our embedding bound holds for the general degree-d case while they only consider the degree-2 case, and even in the degree-2 case, the necessary embedding dimension we provide is $O(\varepsilon^{-2} \cdot \log^3(p))$, which is significantly smaller than the $O(\varepsilon^{-2} \cdot \log^8(p))$ proved in [36].

More peripherally, TensorSketch developed by Pham and Pagh [32] is a popular dimension reduction technique utilizing FFT and fast convolution to recover the Kronecker product of CountSketched [12] vectors. Diao et al. [16] extends the applications of TensorSketch to accelerating Kronecker regression problems by creating oblivious subspace embedding (OSE) [4] without explicitly forming Kronecker products for coefficient matrices.

1.4. Structure of the paper. This paper is developed as follows:

- Section 2 states the main theorem and motivations of our work on Kronecker FJLT;
- Section 3 introduces the technical background and result to support the main theorem;
- Section 4 develops the proof for the technical result;
- Section 5 further discusses on the topic of Kronecker structure's influence to the embedding property and presents related numerical results.

2. Main results of Kronecker FJLT.

The major part of our work is analyzing the vector-based embedding property and providing a theoretical bound of the embedding dimension for Kronecker FJLTs of any degree d.

Theorem 2.1. *Fix $d \geq 1$ and $\varepsilon, \eta \in (0,1)$. Consider integers $n_1, n_2, \ldots, n_d$ and $N = \prod_{k=1}^{d} n_k$. Consider a finite set $\mathcal{E} \subset \mathbb{R}^N$ of cardinality $|\mathcal{E}| = p$. Suppose the Kronecker fast JL transform $\mathbf{\Phi} \in \mathbb{C}^{m \times N}$ has embedding dimension*

$$m \geq C \cdot \left[\varepsilon^{-2} \cdot \log^{2d-1}\left(\frac{p}{\eta}\right) \cdot \log^4\left(\frac{\log(\frac{p}{\eta})}{\varepsilon}\right) \cdot \log N \right]. \tag{2.1}$$

Then with probability at least $1-\eta$, the following holds for all $\mathbf{x} \in \mathcal{E}$:

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2 \tag{2.2}$$

Above, $C > 0$ is an absolute constant.

Remark 2.2. In the case $d = 1$, the embedding $\mathbf{\Phi}$ reduces to the standard fast JL transform corresponding to a random subsampled DFT matrix with randomized column signs. In this case, the results of Theorem 2.1 are already known, see [27], and stated above for completeness. The result for $d \geq 2$ are proved in this paper.

Remark 2.3. In the Kronecker fast JL construction, the randomness in the embedding construction decreases as the degree $d \geq 2$ increases. Specifically, the Kronecker product of independent Rademacher vectors $\bigotimes_{k=d}^{1} \mathbf{D}_{n_k}$ consists of $\sum_{k=1}^{d} n_k$ bits, compared to $N = \prod_{k=1}^{d} n_k$ bits which would be used to construct a standard sign-FJLT. This reduction in randomness is the source of the additional factor of $\log^{2d-2}(p)$ in the number of measurements m required to achieve the quality of approximation compared to the standard FJLT. While we suspect that this additional factor may be pessimistic, some loss of embedding power is necessary with increasing degree d. This is explored numerically in Figure 2.

2.1. Preliminaries. To clearly illustrate our motivation, we first introduce the multilinear algebra background.

Given matrices $\mathbf{A} \in \mathbb{R}^{I\times J}$ and $\mathbf{B} \in \mathbb{R}^{K\times L}$, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is defined as:

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} \mathbf{a}_{11}\mathbf{B} & \mathbf{a}_{12}\mathbf{B} & \cdots & \mathbf{a}_{1J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{I1}\mathbf{B} & \mathbf{a}_{I2}\mathbf{B} & \cdots & \mathbf{a}_{IJ}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{IK\times JL}. \tag{2.3}$$

We will frequently use the distributive property of the Kronecker product in the following development.

$$\mathbf{AB}\otimes\mathbf{CD} = (\mathbf{A}\otimes\mathbf{C})(\mathbf{B}\otimes\mathbf{D}). \tag{2.4}$$

2.2. Cost savings when applied to Kronecker vectors. Although Theorem 2.1 concerns the general embedding property of the KFJLT embedding $\mathbf{\Phi}$, the embedding is particularly useful as an efficient embedding when considered as an operator $\mathbf{\Phi} : \bigotimes_{k=d}^{1}\mathbb{R}^{n_k} \to \mathbb{R}^m$ applied to vectors $\mathbf{x} = \bigotimes_{k=d}^{1}\mathbf{x}_k$ whose with Kronecker product structure matching that of the embedding matrix. In this setting, the Kronecker mixing on $\bigotimes_{k=d}^{1}\mathbf{x}_k$ is equivalent to imposing the mixing operation respectively on each component vector $\mathbf{x}_k$ and most importantly, reduces the mixing cost to a much smaller scale. As the Kronecker structure of the embedded vector is maintained after the mixing, we are able to start from the sampled elements and trace back to find its forming components based on the invertible linear transformation of indices. This strategy restricts the computation objects to only the sampled ones and saves significant amount of floating point operations and memory cost, compared to conventional embedding methods. See Table 1 for the comparison in cost between the standard and Kronecker FJLT on Kronecker vectors.

Table 1
Embedding cost on Kronecker vectors

	Construction	Mixing	Sampling
FJLT	N	$O(N\cdot\log N)+N$	none
KFJLT	none	$O\left(\sum_{k=1}^{d} n_k\log n_k\right)+\sum_{k=1}^{d} n_k$	$d\cdot m$

Note that we treat the construction degree d as a constant in the complexity.

2.3. Applications to CP tensor decomposition. The study of multiway arrays, aka *tensors*, has been an active research area in large-scale data analysis, for its role as a natural algebraic representation for multidimensional data models.

The KFJLT technique is firstly applied as a sketching strategy in a randomized algorithm: CPRAND-MIX for CP tensor decomposition. At each iteration, the *alternating least square* (CP-ALS) problem fitting a rank-r model solves a problem of the form:

$$\min_{\mathbf{X}\in\mathbb{R}^{r\times n}} \|\mathbf{AX}-\mathbf{B}\|_F, \tag{2.5}$$

where

$$\mathbf{A} = (\mathbf{a}(1),\mathbf{a}(2),\cdots,\mathbf{a}(i),\cdots,\mathbf{a}(r)) \in \mathbb{R}^{N\times r} \quad \text{and} \quad \mathbf{B} = (\mathbf{b}(1),\cdots,\mathbf{b}(j),\cdots,\mathbf{b}(n)) \in \mathbb{R}^{N\times n}.$$

Each column $\mathbf{a}(i)$ has the Kronecker structure:

$$\mathbf{a}(i) = \bigotimes_{k=d}^{1} \mathbf{a}_k(i) \in \bigotimes_{k=d}^{1} \mathbb{R}^{n_k} \subset \mathbb{R}^N.$$

This least squares problem is a candidate for the sketching approach $\mathbf{\Phi} = \mathbf{S}\bigotimes_{k=d}^{1}(\mathcal{F}_{n_k}\mathbf{D}_{n_k})$. We refer the readers to Appendix B and [7] for more details.

Theorem 2.1 demonstrates that KFJLT is a low-distortion embedding for a fixed set of points with constant probability. With its application in numerical linear algebra, we can provide a theoretical guarantee for the sample size:

Corollary 2.4. *The KFJLT $\mathbf{\Phi}$ of length N with $O(\varepsilon^{-1}\cdot\sqrt{n}\cdot r^{2d}\cdot\log^{2d+3}(n)\cdot\log N)$ rows is sufficient to output*

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathbb{R}^{r\times n}} \|\mathbf{\Phi}\mathbf{A}\mathbf{X} - \mathbf{\Phi}\mathbf{B}\|_F,$$

such that

$$(1-\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{r\times n}} \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F \le \|\mathbf{A}\hat{\mathbf{X}}-\mathbf{B}\|_F \le (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{r\times n}} \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F \tag{2.6}$$

with high probability.

Proof. Fix $\varepsilon\in(0,1)$ and given a constant failure probability. We first reduce the model to a vector-based problem:

$$\min_{\mathbf{x}\in\mathbb{R}^r} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2, \quad \text{for} \quad \mathbf{b}\in\mathbb{R}^N, \tag{2.7}$$

because the matrix least squares (2.5) on $\mathbf{B}$ can break into the least squares on each column $\mathbf{b}(j)$: $\min_{\mathbf{x}(j)\in\mathbb{R}^r}\|\mathbf{A}\mathbf{x}(j)-\mathbf{b}(j)\|_2$.

Denote $\mathbf{U}$ with $\operatorname{rank}(\mathbf{A}) = r^*$ ($r^*\le r$) columns as the orthogonal basis of $\mathbf{A}$, and let $\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{y}$. The proof of Theorem 2.16 in [38] shows that, if $\mathbf{\Phi}$ satisfies with two conditions: $(1\pm 1/2)$ subspace embedding on the column space of $\mathbf{U}\subset\mathbb{R}^N$, $\mathbf{col}(\mathbf{U})$, and the matrix multiplication

$$\|\mathbf{U}^\top\mathbf{\Phi}^\top\mathbf{\Phi}(\mathbf{U}\hat{\mathbf{y}}-\mathbf{b})\|_2 \le 3\sqrt{\frac{\varepsilon^*}{r}}\cdot\|\mathbf{U}\|_F\cdot\|\mathbf{U}\hat{\mathbf{y}}-\mathbf{b}\|_2 \le 3\sqrt{\varepsilon^*}\cdot\|\mathbf{U}\hat{\mathbf{y}}-\mathbf{b}\|_2, \tag{2.8}$$

then the sketched least squares: $\min_{\mathbf{x}\in\mathbb{R}^r}\|\mathbf{\Phi}\mathbf{A}\mathbf{x}-\mathbf{\Phi}\mathbf{b}\|_2$ can output a $(1\pm\varepsilon^*)$ approximation to the true solution of (2.7). Moreover, (2.8) is achieved from the $(1\pm\sqrt{\varepsilon^*/r})$ embedding on $\mathbf{col}(\mathbf{U})-\mathbf{b}$ and all $\mathbf{U}(i)$, $\mathbf{col}(\mathbf{U})-(\mathbf{b}+\mathbf{U}(i))$, $i\in[r^*]$. Note that $\mathbf{col}(\mathbf{U})-\mathbf{b}$ and $\mathbf{col}(\mathbf{U})-(\mathbf{b}+\mathbf{U}(i))$ can be considered as translations of $\mathbf{col}(\mathbf{U})$, and $\{\mathbf{U}(i)\}_{i\in[r^*]}$ is a subset of r^* vectors.

One can create a $(1\pm\tau)$ subspace embedding for the r^*-dimensional $\mathbf{col}(\mathbf{U})$, and more broadly its translations, by applying the KFJLT $\mathbf{\Phi}$ on a $(1\pm\tau/4)$-approximated net of cardinality on the order of $O(1/\tau)^r$ [21] with distortion less than a multiplicative factor $\tau/2$ [6].

Hence we distribute ε evenly to n vector least squares, $\varepsilon^* = \varepsilon/\sqrt{n}$, and let $\tau = \sqrt{\varepsilon^*/r} = \sqrt{\varepsilon/\sqrt{n}r}$. Recall Theorem 2.1 and apply the result of (2.1): by setting $p = O(\sqrt{\sqrt{n}r/\varepsilon})^r$ and the distortion tolerance $\tau/2 = \sqrt{\varepsilon/4\sqrt{n}r}$ suffices for $\mathbf{\Phi}$ to satisfy the two conditions stated above with a constant probability escape. We obtain the sample size

$$\begin{aligned} &O\left(\frac{\sqrt{n}r}{\varepsilon}\cdot r^{2d-1}\cdot\log^{2d-1}(\sqrt{\frac{\sqrt{n}r}{\varepsilon}})\cdot\log^4(r\cdot\sqrt{\frac{\sqrt{n}r}{\varepsilon}})\cdot\log N\right)\\ &= O\left(\varepsilon^{-1}\cdot\sqrt{n}\cdot r^{2d}\cdot\log^{2d-1}(\frac{\sqrt{n}}{\varepsilon})\cdot\log^4(\frac{\sqrt{n}}{\varepsilon})\cdot\log N\right)\\ &= O\left(\varepsilon^{-1}\cdot\sqrt{n}\cdot r^{2d}\cdot\log^{2d+3}(n)\cdot\log N\right). \end{aligned}$$

We omit factors $1/\varepsilon$ and r inside $\log()$ term by treating ε as a constant and r as a small integer due to the low-rank fitting. We also omit the factor $\log(\sqrt{n}r/\varepsilon)$ in $\log^4()$ compared to the other factor $\sqrt{n}r/\varepsilon$. ■

Notation. $\|\cdot\|_2$ and $\|\cdot\|_\infty$ refer to the ℓ_2 and ℓ_∞ norms of a vector respectively. $\|\cdot\|$, $\|\cdot\|_F$ refer to the spectral and Frobenius norm of a matrix respectively. We use Euler script uppercase letter $\mathcal{X}$ as a tensor, Roman script uppercase letter $\mathbf{X}$ as a matrix, Roman script lowercase letter $\mathbf{x}$ as a vector, simple lowercase letter x as a entry. We put the location index in parentheses $(\cdot)$ and the mode index in subscript $_k$. A capital letter I denotes an index set and a lowercase letter i denotes a singe index. A random vector $\xi \in \mathbb{R}^N$ is said to be a Rademacher sequence if it is uniformly distributed on $\{-1,1\}^N$. $\mathbf{Id}_N \in \mathbb{R}^{N\times N}$ denotes the N by N identity matrix. For a vector $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{D}_\mathbf{x} \in \mathbb{R}^{N\times N}$ denotes the diagonal matrix satisfying $\mathbf{D}_\mathbf{x}(i,i) = \mathbf{x}(i)$ for $i \in [N]$.

3. Proof of Theorem 2.1.

3.1. Background review.

The proof draws on a result established in [27] showing that matrices which can stably embed *sparse* vectors – or have a certain *restricted isometry property* (RIP) [10, 11, 20] – result in Johnson-Lindenstrauss embeddings if their column signs are randomly permuted. First let us recall the definition of the RIP:

Definition 3.1. *A matrix $\mathbf{\Psi} \in \mathbb{R}^{m\times N}$ is said to have the restricted isometry property of order T and level $\delta \in (0,1)$ ((T,δ)-RIP) if*

$$(1-\delta)\|\mathbf{x}\|_2^2 \le \|\mathbf{\Psi}\mathbf{x}\|_2^2 \le (1+\delta)\|\mathbf{x}\|_2^2 \quad \text{for all } T\text{-sparse } \mathbf{x} \in \mathbb{R}^N. \tag{3.1}$$

A vector is T-sparse if it has at most T nonzero entries.

The main result of [27] says that randomizing the columns signs of a (T,δ)-RIP matrix results in a randomized embedding where an arbitrary set of $p = O(e^T)$ points is embedded with multiplicative distortion 4δ, with high probability.

Proposition 3.2 (Theorem 3.1 from [27]). *Fix $\eta > 0$ and $\varepsilon \in (0,1)$ and consider a finite set $\mathcal{E} \subset \mathbb{R}^N$ of cardinality $|\mathcal{E}| = p$. Set $s \ge 20\log(4p/\eta)$ and suppose that $\mathbf{\Psi} \in \mathbb{R}^{m\times N}$ satisfies the restricted isometry property of order $2s$ and level $\delta \le \varepsilon/4$. Let $\xi \in \mathbb{R}^N$ be a Rademacher sequence. Then with probability exceeding $1-\eta$,*

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\mathbf{\Psi}\mathbf{D}_\xi\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \mathcal{E}$.

By the distributive property of the Kronecker product, the KFJLT construction is equivalent to:

$$\mathbf{\Phi} = \mathbf{S}\left(\bigotimes_{k=d}^{1} \mathcal{F}_{n_k}\right)\left(\bigotimes_{k=d}^{1} \mathbf{D}_{n_k}\right).$$

As the Kronecker product preserves the orthogonality, $\bigotimes_{k=d}^{1} \mathcal{F}_{n_k}$ is still a FFT of size N. Thus, we can write equivalently

$$\mathbf{\Phi} = \mathbf{S}\mathcal{F}_N\left(\bigotimes_{k=d}^{1} \mathbf{D}_{\xi_k}\right) = \mathbf{S}\mathcal{F}_N\mathbf{D}_{\xi}. \tag{3.2}$$

where the diagonal entries of each $\mathbf{D}_{\xi_k} \in \mathbb{R}^{n_k \times n_k}$ is built by the corresponding Rademacher sequence $\xi_k \in \mathbb{R}^{n_k}$ and $\bigotimes_{k=d}^{1} \mathbf{D}_{\xi_k}$ is a diagonal matrix depending on the Kronecker product of smaller i.i.d. sequences $\xi = \bigotimes_{k=d}^{1} \xi_k$,

Now, the randomly-subsampled DFT $\mathbf{S}\mathcal{F}_N \in \mathbb{C}^{m \times N}$ is known to satisfy the restricted isometry property with nearly-optimally small embedding dimension m [35, 13, 26, 9, 23]. We state the sharpest known bound, from [23], below.

Proposition 3.3 (Theorem 1.1 from [23]). *For sufficiently large N and T, a unitary matrix $\mathbf{U} \in \mathbb{C}^{N \times N}$ satisfying $\|\mathbf{U}\|_\infty \leq O(1/\sqrt{N})$, and a sufficiently small $\delta > 0$, the following holds. For some $m = O(\log^2(1/\delta) \cdot \delta^{-2} \cdot T \cdot \log^2(T/\delta) \cdot \log N)$, let $\mathbf{\Psi} \in \mathbb{C}^{m \times N}$ be a matrix whose m rows are chosen uniformly and independently from the rows of $\mathbf{U}$, multiplied by N/m. Then, with probability $1 - 2^{-\Omega(\log N \cdot \log(T/\delta))}$, the matrix $\mathbf{\Psi}$ satisfies the restricted isometry property of order T with constant δ.*

Combining this result with Proposition 3.2 proves the embedding result of Theorem 2.1 in the special case $d = 1$. In the case $d \geq 2$, it remains to analyze the effect of applying a Kronecker Rademacher vector to an RIP matrix, as opposed to an i.i.d. Rademacher vector.

3.2. Concentration inequality. We here introduce a more general version of Theorem 2.1, which works for any degree-d construction consisting of a RIP matrix with randomized column sign from a Kronecker product of d independent Rademacher sequences.

Theorem 3.4. *For $d \geq 2$ and given $\varepsilon \in (0, 1)$. $\xi_1 \in \mathbb{R}^{n_1}, \cdots, \xi_d \in \mathbb{R}^{n_d}$ are respectively independent Rademacher sequences. Suppose that $\mathbf{\Psi} \in \mathbb{R}^{m \times N}$ is a $(2s, \delta)$-RIP matrix, where $N = \prod_{k=1}^{d} n_k$. Let $n_1^* \geq n_2^* \geq \cdots \geq n_d^*$ be the decreasing arrangement of the set $\{n_k\}_{k \in [d]}$ and assume that $s \leq n_1^*$. Consider an arbitrary vector $\mathbf{x} \in \mathbb{R}^N$, then*

(3.3)

$$\mathbb{P}\Big(\Big|\|\mathbf{\Psi}\left(\bigotimes_{k=d}^{1}\mathbf{D}_{\xi_k}\right)\mathbf{x}\|_2^2-\|\mathbf{x}\|_2^2\Big|>\varepsilon\cdot\|\mathbf{x}\|_2^2\Big)$$

$$\leq 2\exp(-\frac{1}{88}\cdot\frac{\varepsilon}{\delta\cdot s^{d-2}})\quad+\quad\begin{cases}\underbrace{6(n_2^*)^2}_{C_2}\cdot\exp(-\frac{1}{128}\cdot s):=\boldsymbol{\beta}_2(\varepsilon), & d=2,\\ \underbrace{[6\prod_{k=2}^{d}(n_k^*)^2+2\sum_{\ell=2}^{d-1}\prod_{k=\ell+1}^{d}(n_k^*)^2]}_{C_d}\cdot\exp(-\frac{1}{128}\cdot s):=\boldsymbol{\beta}_d(\varepsilon), & d\geq 3.\end{cases}$$

Remark 3.5. Theorem 2.1 is stated for real-valued embeddings, though the KFJLTs are in the complex field. The result extends to complex matrices straightforwardly via a standard complexification strategy described below. Suppose a partial Fourier matrix $\mathbf{\Psi}=\mathbf{\Psi}_1+\mathbf{i}\cdot\mathbf{\Psi}_2\in\mathbb{C}^{m\times N}$ with $\mathbf{\Psi}_1,\mathbf{\Psi}_2\in\mathbb{R}^{m\times N}$, we map the embedding to a $2m$ tall matrix with $\mathbf{\Psi}_1$ on top and $\mathbf{\Psi}_2$ bottom. The new real-valued matrix satisfies the RIP if $\mathbf{\Psi}$ has this property by equivalence of their operator norms.

$$\|\begin{bmatrix}\mathbf{\Psi}_1\\ \mathbf{\Psi}_2\end{bmatrix}\mathbf{x}\|_2^2=\|\mathbf{\Psi}_1\mathbf{x}\|_2^2+\|\mathbf{\Psi}_2\mathbf{x}\|_2^2=\|\mathbf{\Psi}\mathbf{x}\|_2^2.$$

Rescaling the final result for real-valued embeddings by a factor $1/2$, we obtain the bound of m for the KFJLT constructions.

We now derive Theorem 2.1 from Theorem 3.4, also using the sharpest known RIP bounds on the randomly subsampled DFT matrix Proposition 3.3.

Proof of Theorem 2.1 for the $d\geq 2$ case. Recall the degree-d KFJLT construction $\mathbf{\Phi}=\mathbf{S}\mathcal{F}_N\bigotimes_{k=d}^{1}\mathbf{D}_{\xi_k}\in\mathbb{C}^{m\times N}$, where N is the product of each mode size n_k. Given a fixed distortion tolerance $\varepsilon\in(0,1)$, we focus on the complement event $\mathbf{F}$ of (2.2) achieved for all p points in $\mathcal{E}\subset\mathbb{R}^N$: there exists vector $\mathbf{x}\in\mathcal{E}$ such that $|\|\mathbf{\Phi}\mathbf{x}\|_2^2-\|\mathbf{x}\|_2^2|>\varepsilon\|\mathbf{x}\|_2^2$. As the uniformly sampled $\mathbf{S}\mathcal{F}_N\in\mathbb{C}^{m\times N}$ is a RIP matrix $\mathbf{\Psi}$, suppose that it satisfies the recovering level δ and order $T=2s\leq 2\max_{k\in[d]}n_k$. Take a union of $\mathbf{\Phi}$ failing to embed a single vector described in Theorem 3.4 on the entire $\mathcal{E}$, by multiplying p with the result from (3.3), we obtain an upper bound for $\mathbb{P}(\mathbf{F})$:

$$\underbrace{2p\cdot\exp(-\frac{1}{88}\cdot\frac{\varepsilon}{\delta\cdot s^{d-2}})}_{(\mathbf{I})}+\underbrace{2p\cdot(d+1)\cdot N^{2-\frac{2}{d}}\cdot\exp(-\frac{1}{128}\cdot s)}_{(\mathbf{II})}.$$

Note that $(\mathbf{II})$ is simplified from (3.3) due to $C_d\leq(6+2(d-2))\cdot(N/n_1^*)^2$ and $n_1^*\geq N^{1/d}$ when $d\geq 2$.

We aim to restrict $\mathbb{P}(\mathbf{F})$ within a small $\eta>0$ in order for the $(1\pm\varepsilon)$ embedding on $\mathcal{E}$ to happen with high probability. Hence by bounding $(\mathbf{I}),(\mathbf{II})$ respectively in $\eta/2,\eta/2$, it leaves

conditions on the RIP variables:

$$\begin{cases} n_1^* \geq s \geq 128 \log\left((d+1)\cdot N^{2-\frac{2}{d}}\cdot \dfrac{4p}{\eta}\right), \\ \delta \cdot s^{d-2} \leq \dfrac{\varepsilon}{88\log(\frac{4p}{\eta})}. \end{cases} \tag{3.4}$$

Without loss, suppose $p > N$, because we can always embed an arbitrary set of vectors of cardinality less than N from $\mathbb{R}^N$ into its subspace. Hence $s = O(\log(d \cdot N^{2-\frac{2}{d}} \cdot p/\eta))$ is on the same order as $\log(p/\eta)$, if treating the degree d as a constant in the complexity. Directly from the second condition in (3.4), $\delta = o(\varepsilon/(s^{d-2} \cdot \log(p/\eta)))$, which is restricted on the order of $\log(p/\eta)$ to the power $d-1$ dividing ε.

The major multiplicative factor of the number of measurements presented in Proposition 3.3 is $\delta^{-2} \cdot s$. From the discussion above, for $d \geq 2$, we obtain a scale of m depending on d as the power number which cannot be omitted:

$$\varepsilon^{-2} \cdot \log^{2d-2}(\frac{p}{\eta}) \cdot \log(\frac{p}{\eta}) = \varepsilon^{-2} \cdot \log^{2d-1}(\frac{p}{\eta}). \tag{3.5}$$

More specifically,

$$\begin{aligned} m &\gtrsim \delta^{-2} \cdot s \cdot \log^2(\frac{1}{\delta}) \cdot \log^2(\frac{s}{\delta}) \cdot \log N \\ &\gtrsim \varepsilon^{-2} \cdot \log^{2d-1}(\frac{p}{\eta}) \cdot \log^2(\frac{\log^{d-1}(\frac{p}{\eta})}{\varepsilon}) \cdot \log^2(\frac{\log^{d}(\frac{p}{\eta})}{\varepsilon}) \cdot \log N \\ &\gtrsim \varepsilon^{-2} \cdot \log^{2d-1}(\frac{p}{\eta}) \cdot \log^4(\frac{\log^{d}(\frac{p}{\eta})}{\varepsilon}) \cdot \log N \\ &= O\left(\varepsilon^{-2} \cdot \log^{2d-1}(\frac{p}{\eta}) \cdot \log^4(\frac{\log(\frac{p}{\eta})}{\varepsilon}) \cdot \log N\right). \end{aligned}$$

■

Remark 3.6. Be aware that Proposition 3.3 shows $\mathbf{\Psi}$ satisfies the RIP except for a small probability $2^{-\Omega(\log N \cdot \log(T/\delta))}$. However, given ε and η are constants in the JL result Theorem 3.4, T/δ is on the order of $\log^d(p)$, thus

$$2^{-\Omega(\log N \cdot \log(T/\delta))} \lesssim (\frac{1}{\log p})^{d \cdot \log N},$$

which is significantly small compared to a constant η, as p is usually exponentially large.

4. Proof of Theorem 3.4.

4.1. Proof ingredients. We recall basic corollaries of the restricted isometry property, whose proofs can be found in [33]. Suppose that $\mathbf{\Psi} \in \mathbb{R}^{m \times n}$ has the restricted isometry property of order $2s$ and level δ, for an arbitrary vector $\mathbf{x} \in \mathbb{R}^n$. Then

1. Lemma 4.1. *For a subset $I \subset [n]$ of size $|I| \leq s$,*

$$\left| \|\mathbf{\Psi}(I)\mathbf{x}(I)\|_2^2 - \|\mathbf{x}(I)\|_2^2 \right| \leq \delta \cdot \|\mathbf{x}(I)\|_2^2. \tag{4.1}$$

2. Lemma 4.2. *For any pair of disjoint subsets $I, J \subset [n]$ of size $|I|, |J| \leq s$,*

$$\left|\langle \mathbf{\Psi}(I)\mathbf{x}(I), \mathbf{\Psi}(J)\mathbf{x}(J)\rangle\right| \leq \delta \cdot \|\mathbf{x}(I)\|_2 \cdot \|\mathbf{x}(J)\|_2. \tag{4.2}$$

Then let us recall standard concentration inequalities in both linear and quadratic forms, particularly for Rademacher sequences:

Lemma 4.3 (Hoeffding's inequality). *Let $\mathbf{x} \in \mathbb{R}^n$ be a sequence and $\xi \in \mathbb{R}^n$ be a Rademacher sequence. Then, for any $t > 0$,*

$$\mathbb{P}(|\xi^\top \mathbf{x}| > t) \leq 2\exp\left(-\frac{t^2}{2\|\mathbf{x}\|_2^2}\right). \tag{4.3}$$

This version of Hoeffding's inequality is derived directly from Theorem 2 of [24].

Lemma 4.4 (Hanson-Wright inequality). *[22] Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ have zero diagonal entries, and $\xi \in \mathbb{R}^n$ be a Rademacher sequence. Then, for any $t > 0$,*

$$\mathbb{P}(|\xi^\top \mathbf{X}\xi| > t) \leq 2\exp\left(-\frac{1}{64}\min\left(\frac{t^2}{\|\mathbf{X}\|_F^2}, \frac{\frac{96}{65}t}{\|\mathbf{X}\|_2}\right)\right). \tag{4.4}$$

This Hanson-Wright bound with explicit constants is derived from the proof of Theorem 17 in [8].

We will use the following corollary of Hanson-Wright.

Corollary 4.5. *Suppose we have a random matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, positive vectors $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$, and $\tau > 0$, $\boldsymbol{\beta} > 0$ such that, for each pair $(i, j) \in [n]$,*

$$\mathbb{P}\left(|x(i,j)| > \tau \cdot \mathbf{y}_1(i) \cdot \mathbf{y}_2(j)\right) \leq \boldsymbol{\beta}.$$

Then for an independent Rademacher sequence $\xi \in \mathbb{R}^n$, and $t > 0$ such that $\tau \leq t/66$, we have

$$\mathbb{P}\left(|\xi^\top \mathbf{X}\xi| > t \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2\right) \leq n^2 \cdot \boldsymbol{\beta} + 2\exp\left(-\frac{1}{44} \cdot \frac{t}{\tau}\right), \tag{4.5}$$

where the probability is with respect to both $\mathbf{X}$ and ξ.

Proof. With probability at least $1 - n^2 \cdot \boldsymbol{\beta}$ with respect to the draw of $\mathbf{X}$, $\{|x(i,j)| \leq \tau \cdot \mathbf{y}(i) \cdot \mathbf{y}(j)\}$ for all $i, j \in [n]$.

By the law of total probability,

$$\begin{aligned}
&\mathbb{P}\left(|\xi^\top \mathbf{X}\xi| > t \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2\right) \\
&\leq n^2 \cdot \boldsymbol{\beta} + \mathbb{P}\left(|\xi^\top \mathbf{X}\xi| > t \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2 \;\middle|\; \{|x(i,j)| \leq \tau \cdot \mathbf{y}_1(i) \cdot \mathbf{y}_2(j)\} \text{ for all } i, j \in [n]\right) \\
&\leq n^2 \cdot \boldsymbol{\beta} + \mathbb{P}\left(|\xi^\top \mathbf{X}\xi| > t \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2 \;\middle|\; \mathbf{E}\right),
\end{aligned}$$

where $\mathbf{E}$ is the event

$$\mathbf{E} = \{|\mathrm{Tr}(\mathbf{X})| \leq \tau \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2, \quad \|\tilde{\mathbf{X}}\| \leq \tau \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2, \quad \|\tilde{\mathbf{X}}\|_F \leq \tau \cdot \|\mathbf{y}_1\|_2 \cdot \|\mathbf{y}_2\|_2\},$$

and $\mathrm{Tr}(\mathbf{X})$ is the trace of $\mathbf{X}$, and $\tilde{\mathbf{X}}$ is formed from $\mathbf{X}$ by setting the diagonal entries to zero.

Then, applying Hanson-Wright,

$$\begin{aligned}
&\mathbb{P}\left(\left|\xi^\top \mathbf{X}\xi\right| > t\cdot\|\mathbf{y}_1\|_2\cdot\|\mathbf{y}_2\|_2 \;\middle|\; \mathbf{E}\right)\\
&\le \mathbb{P}\left(|\mathrm{Tr}(\mathbf{X})| > \tau\cdot\|\mathbf{y}_1\|_2\cdot\|\mathbf{y}_2\|_2 \;\middle|\; \mathbf{E}\right) + \mathbb{P}\left(\left|\xi^\top \tilde{\mathbf{X}}\xi\right| > (t-\tau)\cdot\|\mathbf{y}_1\|_2\cdot\|\mathbf{y}_2\|_2 \;\middle|\; \mathbf{E}\right)\\
&= \mathbb{P}\left(\left|\xi^\top \tilde{\mathbf{X}}\xi\right| > (t-\tau)\cdot\|\mathbf{y}_1\|_2\cdot\|\mathbf{y}_2\|_2 \;\middle|\; \mathbf{E}\right)\\
&\le 2\exp\left(-\min(\frac{1}{64}\cdot\frac{(t-\tau)^2}{\tau^2}, \frac{3}{130}\cdot\frac{t-\tau}{\tau})\right) = 2\exp(-\frac{3}{130}\cdot\frac{t-\tau}{\tau})\\
&\le 2\exp(-\frac{1}{44}\cdot\frac{t}{\tau}) \quad (\text{if } \tau \le t/66).
\end{aligned}$$

Therefore we obtain a upper bound: $\mathbb{P}\left(\left|\xi^\top \mathbf{X}\xi\right| > t\cdot\|\mathbf{y}_1\|_2\cdot\|\mathbf{y}_2\|_2\right) \le n^2\cdot\boldsymbol{\beta} + 2\exp(-t/44\tau)$. ■

The following proposition is similar to Proposition 5.4 in [27], adapted to apply to general quadratic forms as opposed to symmetric ones.

Proposition 4.6. *Fix integers s, n, m such that $s \le n$ and let $r = \lceil n/s \rceil$. Let $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_L, \boldsymbol{\Psi}_R) \in \mathbb{R}^{m\times 2n}$, where $\boldsymbol{\Psi}_L, \boldsymbol{\Psi}_R \in \mathbb{R}^{m\times n}$ respectively denote the first and the second sets of n columns, have the $(2s,\delta)$-RIP. Consider arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Let I_1 of size s be the index set containing the largest s-magnitude entries of $\mathbf{x}$, I_2 the index set containing the largest s-magnitude (possibly less than s) entries among the entries indexed by I_1^c, up to I_r. The corresponding index notations for $\mathbf{y}$ are the sets $J_1, \cdots, J_r$. We write $i_1 \sim j_1$ if the two indices are associated in the same block location respectively of $\mathbf{x}$ and $\mathbf{y}$, i.e. $i_1 \in I_p, j_1 \in J_p$, $p = 1, \cdots, r$. Consider the matrix $\mathbf{C}_{\mathbf{x},\mathbf{y}} \in \mathbb{R}^{n\times n}$ with entries:*

$$\mathbf{C}_{\mathbf{x},\mathbf{y}}(i,j) = \begin{cases} x(i)\boldsymbol{\Psi}_L(i)^\top\boldsymbol{\Psi}_R(j)y(j), & i \not\sim j, \quad i \in I_1^c, j \in J_1^c,\\ 0, & \textit{else.}\end{cases}$$

And for $\mathbf{b} \in \{-1,1\}^s$, $\mathbf{d} \in \{-1,1\}^n$,

$$\begin{aligned}
\mathbf{v}_{\mathbf{x},\mathbf{y}} &= \mathbf{D}_{\mathbf{x}(I_1^c)}\boldsymbol{\Psi}_L(I_1^c)^\top\boldsymbol{\Psi}_R(J_1)\mathbf{D}_{\mathbf{y}(J_1)}\mathbf{b} \in \mathbb{R}^{n-s},\\
\mathbf{W}_{\mathbf{x},\mathbf{y}} &= \sum_{p=1}^{r}\mathbf{d}(I_p)^\top\mathbf{D}_{\mathbf{x}(I_p)}\boldsymbol{\Psi}_L(I_p)^\top\boldsymbol{\Psi}_R(J_p)\mathbf{D}_{\mathbf{y}(J_p)}\mathbf{d}(J_p) \in \mathbb{R}.
\end{aligned}$$

Then

$$\begin{aligned}
\|\mathbf{C}_{\mathbf{x},\mathbf{y}}\| &\le \frac{\delta}{s}\cdot\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2, \qquad \|\mathbf{C}_{\mathbf{x},\mathbf{y}}\|_F \le \frac{\delta}{\sqrt{s}}\cdot\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2,\\
\|\mathbf{v}_{\mathbf{x},\mathbf{y}}\|_2 &\le \frac{\delta}{\sqrt{s}}\cdot\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2, \qquad |\mathbf{W}_{\mathbf{x},\mathbf{y}}| \le \delta\cdot\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2.
\end{aligned}$$

The detailed proof of Proposition 4.6 can be found in Appendix A.

4.2. Notations. Without loss of generality, assume that $n_1 \geq n_2 \geq \cdots \geq n_d$, because we can always make the mode size in a decreasing arrangement by permuting the corresponding columns.

For $d \geq 2$, we analyze the JL construction in a block manner. Consider the block decomposition:

$$\begin{aligned} \mathbf{x} &= (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_{n_d}) \in \mathbb{R}^{\prod_{k=1}^{d} n_k}, \\ \mathbf{\Psi} &= (\mathbf{\Psi}_1, \mathbf{\Psi}_2, \cdots, \mathbf{\Psi}_i, \cdots, \mathbf{\Psi}_{n_d}) \in \mathbb{R}^{m \times \prod_{k=1}^{d} n_k}, \end{aligned} \tag{4.6}$$

where the block vector $\mathbf{x}_i \in \mathbb{R}^{\prod_{k=1}^{d-1} n_k}$ and the block matrix $\mathbf{\Psi}_i \in \mathbb{R}^{m \times \prod_{k=1}^{d-1} n_k}$ respectively denote the entries of $\mathbf{x}$ and the columns of $\mathbf{\Psi}$ indexed in $\{(i-1)\cdot\prod_{k=1}^{d-1} n_k + 1, \cdots, i\cdot\prod_{k=1}^{d-1} n_k\}$.

Each of the blocks $\{\mathbf{\Psi}_i\}_{i\in[n_d]}$ has the $(2s,\delta)$-RIP when $d \geq 2$ as $\mathbf{\Psi}$ does and s is less than the first mode size n_1. Therefore

$$\mathbf{\Phi}_i = \mathbf{\Psi}_i\Big(\bigotimes_{k=d-1}^{1} \mathbf{D}_{\xi_k}\Big) \in \mathbb{R}^{m \times \prod_{k=1}^{d-1} n_k} \tag{4.7}$$

is a degree-$(d-1)$ JL construction. In the column arrangement, $\mathbf{\Phi}$ consists of a series of blocks $\mathbf{\Phi}_i$, with randomized signs determined by the random sequence ξ_d:

$$\mathbf{\Phi} = (\xi_d(1)\cdot\mathbf{\Phi}_1, \cdots, \xi_d(i)\cdot\mathbf{\Phi}_i, \cdots, \xi_d(n_d)\cdot\mathbf{\Phi}_{n_d}) \in \mathbb{R}^{m \times \prod_{k=1}^{d} n_k}. \tag{4.8}$$

Now we focus on the distortion

$$\|\mathbf{\Phi}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 = \|\mathbf{\Psi}\mathbf{D}_\xi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \tag{4.9}$$

and start by writing it as a quadratic form:

$$\xi^\top \mathbf{M} \xi,$$

where $\mathbf{M} = \mathbf{D}_\mathbf{x}(\mathbf{\Psi}^\top\mathbf{\Psi} - \mathbf{Id})\mathbf{D}_\mathbf{x} \in \mathbb{R}^{\prod_{k=1}^{d} n_k \times \prod_{k=1}^{d} n_k}$.

In fact, based on the block decomposition in (4.8), we can write the distortion also as

$$\xi_d^\top \mathbf{M}_d \xi_d, \tag{4.10}$$

where $\mathbf{M}_d \in \mathbb{R}^{n_d \times n_d}$ has entries

$$m_d(i,j) = \begin{cases} \|\mathbf{\Phi}_i\mathbf{x}_i\|_2^2 - \|\mathbf{x}_i\|_2^2, & i = j \\ \langle \mathbf{\Phi}_i\mathbf{x}_i, \mathbf{\Phi}_j\mathbf{x}_j\rangle, & i \neq j \end{cases}. \tag{4.11}$$

4.3. The main proof. We use the induction method to prove Theorem 3.4 for a general d. The proof proceeds by proving the following result by induction on degree d. Note that the degree $d-1$ case of Theorem 3.4 follows by applying the following proposition to the case d and $i = j$.

Proposition 4.7. *Given a distortion tolerance $\varepsilon_d \in (0,1)$ for all $d \geq 2$, let $\mathbf{\Psi} \in \mathbb{R}^{m \times \prod_{k=1}^{d} n_k}$ be partitioned in the way of (4.6), where $\mathbf{\Psi}_i \in \mathbb{R}^{m \times \prod_{k=1}^{d-1} n_k}$ denotes a block matrix, and $\xi_1 \in \mathbb{R}^{n_1}, \cdots, \xi_d \in \mathbb{R}^{n_d}$ are respectively independent Rademacher sequences. Suppose that $s \leq n_1$ and $\mathbf{\Psi}$ has the $(2s,\delta)$-RIP. Consider arbitrary vectors $\{\mathbf{x}_i\}_{i\in[n_d]} \in \mathbb{R}^{\prod_{k=1}^{d-1} n_k}$, following the notations in (4.7) and (4.11), then it holds for each pair $(i,j) \in [n_d] \times [n_d]$,*

$$\mathbb{P}\Big(|m_d(i,j)| > \varepsilon_d \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2\Big)$$

$$\leq \begin{cases} 6\exp(-\dfrac{1}{128}\cdot s) := \boldsymbol{\beta}_1(2\delta), & d=2 \\ 2\exp(-\dfrac{1}{88}\cdot\dfrac{\varepsilon_3}{\delta}) + 6n_2^2\cdot\exp(-\dfrac{1}{128}\cdot s) := \boldsymbol{\beta}_2(\varepsilon_3), & d=3 \\ 2\exp(-\dfrac{1}{88}\cdot\dfrac{\varepsilon_d}{\delta\cdot s^{d-3}}) + (6\prod_{k=2}^{d} n_k^2 + 2\sum_{\ell=2}^{d-1}\prod_{k=\ell+1}^{d} n_k^2)\cdot\exp(-\dfrac{1}{128}\cdot s) := \boldsymbol{\beta}_{d-1}(\varepsilon_d), & d\geq 4, \end{cases},$$

with $\varepsilon_2 = 2\delta$ pre-set for the base case.

Now we show the proof of Proposition 4.7 by induction for $d \geq 2$.

Proof. 1. The base case $d = 2$.

For the base case, we employ a similar technique as in the proof of Proposition 3.2 in [27] for the standard FJLT. Recall the matrix m_2 of dimension $[n_1] \times [n_2]$ from (4.11) with entries

$$m_2(i,j) = \xi_1^\top \mathbf{M}_{i,j}\xi_1, \tag{4.12}$$

where $\mathbf{M}_{i,j} \in \mathbb{R}^{n_1\times n_1}$:

$$\begin{cases} \mathbf{D}_{\mathbf{x}_i}(\mathbf{\Psi}_i^\top\mathbf{\Psi}_i - \mathbf{Id}_{n_1})\mathbf{D}_{\mathbf{x}_i}, & i=j, \\ \mathbf{D}_{\mathbf{x}_i}\mathbf{\Psi}_i^\top\mathbf{\Psi}_j\mathbf{D}_{\mathbf{x}_j}, & i\neq j. \end{cases}$$

Consider the matrix $\mathbf{C}_{i,j} \in \mathbb{R}^{n_1\times n_1}$ with entries:

$$\mathbf{C}_{i,j}(i_1,j_1) = \begin{cases} m_{i,j}(i_1,j_1), & i_1 \not\sim j_1, \quad i_1 \in I_1^c, j_1 \in J_1^c, \\ 0, & \text{else.} \end{cases}$$

And,

$$\mathbf{v}_{i,j} = \mathbf{M}_{i,j}(I_1^c, J_1)\xi_1(J_1) \in \mathbb{R}^{n_1-s},$$

$$\mathbf{W}_{i,j} = \sum_{p=1}^{r} \xi_1(I_p)^\top \mathbf{M}_{i,j}(I_p,J_p)\xi_1(J_p) \in \mathbb{R}.$$

By directly applying the result from Proposition 5.4 in [27] if $i = j$, and Proposition 4.6 if $i \neq j$, the norm bounds hold:

$$\|\mathbf{C}_{i,j}\| \leq \frac{\delta}{s} \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2, \quad \|\mathbf{C}_{i,j}\|_F \leq \frac{\delta}{\sqrt{s}} \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2,$$
$$\|\mathbf{v}_{i,j}\|_2 \leq \frac{\delta}{\sqrt{s}} \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2, \quad |\mathbf{W}_{i,j}| \leq \delta \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2.$$

Moreover,

$$\begin{aligned} m_2(i,j) = \xi_1^\top \mathbf{M}_{i,j} \xi_1 &= \sum_{p,q=1}^{r} \xi_1(I_p)^\top \mathbf{M}_{i,j}(I_p, J_q) \xi_1(J_q) \\ &= \xi_1^\top \mathbf{C}_{i,j} \xi_1 + \xi_1(I_1^c)^\top \mathbf{v}_{i,j} + \xi_1(J_1^c)^\top \mathbf{v}_{j,i} + \mathbf{W}_{i,j}, \end{aligned}$$

since

$$\xi_1^\top \mathbf{C}_{i,j} \xi_1 = \sum_{\substack{p,q=2 \\ p \neq q}}^{r} \xi_1(I_p)^\top \mathbf{M}_{i,j}(I_p, J_q) \xi_1(J_q),$$
$$\xi_1(I_1^c)^\top \mathbf{v}_{i,j} = \sum_{p=2}^{r} \xi_1(I_p)^\top \mathbf{M}_{i,j}(I_p, J_1) \xi_1(J_1),$$
$$\mathbf{v}_{j,i}^\top \xi_1(J_1^c) = \sum_{q=2}^{r} \xi_1(I_1)^\top \mathbf{M}_{i,j}(I_1, J_q) \xi_1(J_q) = \xi_1(J_1^c)^\top \mathbf{v}_{j,i}.$$

By the standard concentration inequalities Lemma 4.3 and Lemma 4.4,

$$\begin{aligned} \mathbb{P}\left(|\xi_1(I_1^c)^\top \mathbf{v}_{i,j}| > \frac{1}{8} \cdot \delta \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2 \right) &\leq 2 \exp\left(-\frac{1}{2} \cdot \frac{\delta^2}{64} \cdot \frac{s}{\delta^2} \right) \\ &= 2 \exp(-\frac{1}{128} \cdot s), \end{aligned}$$

$$\mathbb{P}\left(|\xi_1(J_1^c)^\top \mathbf{v}_{j,i}| > \frac{1}{8} \cdot \delta \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2 \right) \leq 2 \exp(-\frac{1}{128} \cdot s),$$

$$\begin{aligned} \mathbb{P}\left(|\xi_1^\top \mathbf{C}_{i,j} \xi_1| > \frac{3}{4} \cdot \delta \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2 \right) &\leq 2 \exp\left(-\min(\frac{1}{64} \cdot \frac{9\delta^2}{16} \cdot \frac{s}{\delta^2}, \frac{3}{130} \cdot \frac{3\delta}{4} \cdot \frac{s}{\delta}) \right) \\ &= 2 \exp\left(-\min(\frac{9}{1024} \cdot s, \frac{9}{520} \cdot s) \right) \\ &< 2 \exp(-\frac{1}{128} \cdot s). \end{aligned}$$

Note that as I_1 and J_1 are chosen independently when $i \neq j$, $\xi_1(I_1^c)^\top \mathbf{v}_{i,j}$ and $\xi_1(J_1^c)^\top \mathbf{v}_{j,i}$ are estimated separately.

Finally, as

$$\begin{cases} |\xi_1(I_1^c)^\top \mathbf{v}_{i,j}| > \dfrac{1}{8}\cdot\delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2, \\ |\xi_1(J_1^c)^\top \mathbf{v}_{j,i}| > \dfrac{1}{8}\cdot\delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2, \\ |\xi_1^\top \mathbf{C}_{i,j}\xi_1| > \dfrac{3}{4}\cdot\delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2, \\ |\mathbf{W}_{i,j}| > \delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2, \end{cases}$$

imply $|m_2(i,j)| > 2\delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2$. By the law of total probability, we obtain a uniform bound for each pair $(i,j)\in[n_2]\times[n_2]$,

$$\begin{aligned}\mathbb{P}\left(\left|m_2(i,j)\right| > 2\delta\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2\right) &\le 2\exp(-\frac{1}{128}\cdot s)+2\exp(-\frac{1}{128}\cdot s)+2\exp(-\frac{1}{128}\cdot s)\\ &\le 6\exp(-\frac{1}{128}\cdot s) := \boldsymbol{\beta}_1(2\delta).\end{aligned}$$

.

2. The induction step.

   Suppose now that Proposition 4.7 is true up to degree $d\ge 2$. We aim to show that the statement must then hold also for degree $d+1$.

   In the degree-$(d+1)$ case, for $i\in[n_{d+1}]$, consider one further step of the block decomposition:

$$\begin{aligned}\mathbf{x}_i &= \left(\mathbf{x}_{i,1},\mathbf{x}_{i,2},\cdots,\mathbf{x}_{i,i'},\cdots,\mathbf{x}_{i,n_d}\right)\in\mathbb{R}^{\prod_{k=1}^d n_k},\\ \boldsymbol{\Psi}_i &= \left(\boldsymbol{\Psi}_{i,1},\boldsymbol{\Psi}_{i,2},\cdots,\boldsymbol{\Psi}_{i,i'},\cdots,\boldsymbol{\Psi}_{i,n_d}\right)\in\mathbb{R}^{m\times\prod_{k=1}^d n_k},\end{aligned}$$

   where the block vector $\mathbf{x}_{i,i'}\in\mathbb{R}^{\prod_{k=1}^{d-1} n_k}$ and the block matrix $\boldsymbol{\Psi}_{i,i'}\in\mathbb{R}^{m\times\prod_{k=1}^{d-1} n_k}$ respectively denote the entries of $\mathbf{x}$ and the columns of $\boldsymbol{\Psi}$ indexed in the set $\{(i-1)\cdot\prod_{k=1}^d n_k+(i'-1)\cdot\prod_{k=1}^{d-1} n_k+1,\cdots,(i-1)\cdot\prod_{k=1}^d n_k+i'\cdot\prod_{k=1}^{d-1} n_k\}$. The corresponding construction:

$$\boldsymbol{\Phi}_{i,i'} = \boldsymbol{\Psi}_{i,i'}\Big(\bigotimes_{k=d-1}^{1}\mathbf{D}_{\xi_k}\Big)$$

   is a degree-$(d-1)$ JL embedding.

   Recall the form (4.10) of m_{d+1}. For each pair of $(i,j)\in[n_{d+1}]\times[n_{d+1}]$,

$$m_{d+1}(i,j) = \xi_d^\top(\mathbf{M}_d)_{i,j}\xi_d,$$

   where the center matrix $(\mathbf{M}_d)_{i,j}\in\mathbb{R}^{n_d\times n_d}$ has entries:

$$(m_d)_{i,j}(i',j') = \begin{cases}\|\boldsymbol{\Phi}_{i,i'}\mathbf{x}_{i,i'}\|_2^2-\|\mathbf{x}_{i,i'}\|_2^2, & \text{if } i=j \text{ and } i'=j',\\ \langle\boldsymbol{\Phi}_{i,i'}\mathbf{x}_{i,i'},\boldsymbol{\Phi}_{j,j'}\mathbf{x}_{j,j'}\rangle, & \text{else.}\end{cases}$$

are in the category of entries m_d showed in (4.11) since each $\mathbf{\Phi}_{i,i'}$ is a $(2s,\delta)$-RIP $\mathbf{\Psi}_{i,i'}$ with randomized column signs from $\bigotimes_{k=d-1}^{1}\xi_k$ and the index sets of $\mathbf{\Phi}_{i,i'}, \mathbf{\Phi}_{j,j'}$ are disjoint except when both $i=j$ and $i'=j'$.

Following the eligibility of applying the result for m_d in the center matrix of $m_{d+1}(i,j)$, in particular, set specifically $\varepsilon_d = 2\delta\cdot s^{d-2} \in (0,1)$, such that for each pair $(i',j') \in [n_d]\times[n_d]$,

$$
\begin{aligned}
&\mathbb{P}\Big(|(m_d)_{i,j}(i',j')| > 2\delta\cdot s^{d-2}\cdot\|\mathbf{x}_{i,i'}\|_2\cdot\|\mathbf{x}_{j,j'}\|_2\Big)\\
&\leq \begin{cases} 6\exp(-\dfrac{1}{128}\cdot s), & d=2\\ 6n_2^2\cdot\exp(-\dfrac{1}{128}\cdot s)+2\exp(-\dfrac{1}{44}\cdot s), & d=3\\ (6\displaystyle\prod_{k=2}^{d}n_k^2+2\sum_{\ell=2}^{d-1}\prod_{k=\ell+1}^{d}n_k^2)\cdot\exp(-\dfrac{1}{128}\cdot s)+2\exp(-\dfrac{1}{44}\cdot s), & d\geq 4\end{cases}.
\end{aligned}
\tag{4.13}
$$

Given $\varepsilon_{d+1}\in(0,1)$ and apply Corollary 4.5,

$$
\begin{aligned}
&\mathbb{P}\Big(|m_3(i,j)| > \varepsilon_3\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2\Big)\\
&\leq 2\exp(-\frac{1}{44}\cdot\frac{\varepsilon_3}{2\delta})+n_2^2\cdot 6\exp(-\frac{1}{128}\cdot s)\\
&\leq 2\exp(-\frac{1}{88}\cdot\frac{\varepsilon_3}{\delta})+6n_2^2\cdot\exp(-\frac{1}{128}\cdot s) := \boldsymbol{\beta}_2(\varepsilon_3).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\mathbb{P}\Big(|m_4(i,j)| > \varepsilon_4\cdot\|\mathbf{x}_i\|_2\cdot\|\mathbf{x}_j\|_2\Big)\\
&\leq 2\exp(-\frac{1}{44}\cdot\frac{\varepsilon_4}{2\delta\cdot s})+n_3^2\cdot[6n_2^2\cdot\exp(-\frac{1}{128}\cdot s)+2\exp(-\frac{1}{44}\cdot s)]\\
&\leq 2\exp(-\frac{1}{88}\cdot\frac{\varepsilon_4}{\delta\cdot s})+(6n_3^2n_2^2+2n_3^2)\cdot\exp(-\frac{1}{128}\cdot s) := \boldsymbol{\beta}_3(\varepsilon_4).
\end{aligned}
$$

Note that we simply replace $\exp(-s/44)$ with $\exp(-s/128)$ in the last step derivation as the latter of bigger value works for an upper bound to make a more organized result.

For general case $d \geq 4$,

$$\mathbb{P}\Big(|m_{d+1}(i,j)| > \varepsilon_{d+1} \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2\Big)$$

$$\leq 2\exp(-\frac{1}{44} \cdot \frac{\varepsilon}{2\delta \cdot s^{d-2}}) + n_d^2 \cdot [(6\prod_{k=2}^{d-1} n_k^2 + 2\sum_{\ell=2}^{d-2}\prod_{k=\ell+1}^{d-1} n_k^2) \cdot \exp(-\frac{1}{128} \cdot s) + 2\exp(-\frac{1}{44} \cdot s)]$$

$$\leq 2\exp(-\frac{1}{88} \cdot \frac{\varepsilon_{d+1}}{\delta \cdot s^{d-2}}) + (6\prod_{k=2}^{d} n_k^2 + 2\sum_{\ell=2}^{d-1}\prod_{k=\ell+1}^{d} n_k^2) \cdot \exp(-\frac{1}{128} \cdot s) := \boldsymbol{\beta}_d(\varepsilon_{d+1}).$$

Also note that we assume $2\delta \cdot s^{d-2} \leq \varepsilon_{d+1}/66$ in order to apply Corollary 4.5. Our final constraint on δ, s and the distortion tolerance showed in (3.4): $\delta \cdot s^{d-2} \leq \varepsilon/(88\log(4p/\eta))$ makes the assumption valid as the logarithm is usually considered greater than 2.

We finally obtain the result for $d+1$ and complete the proof for Proposition 4.7. ■

5. Numerical experiments and further discussions. In this section, we run numerical experiments to study the empirical embedding performance of Kronecker FJLT. It is of value to discuss and compare the performance of KFJLT with varying degree d, including the standard FJLT corresponding to $d = 1$, in order to evaluate the trade-off between distortion power and computational speed-up.

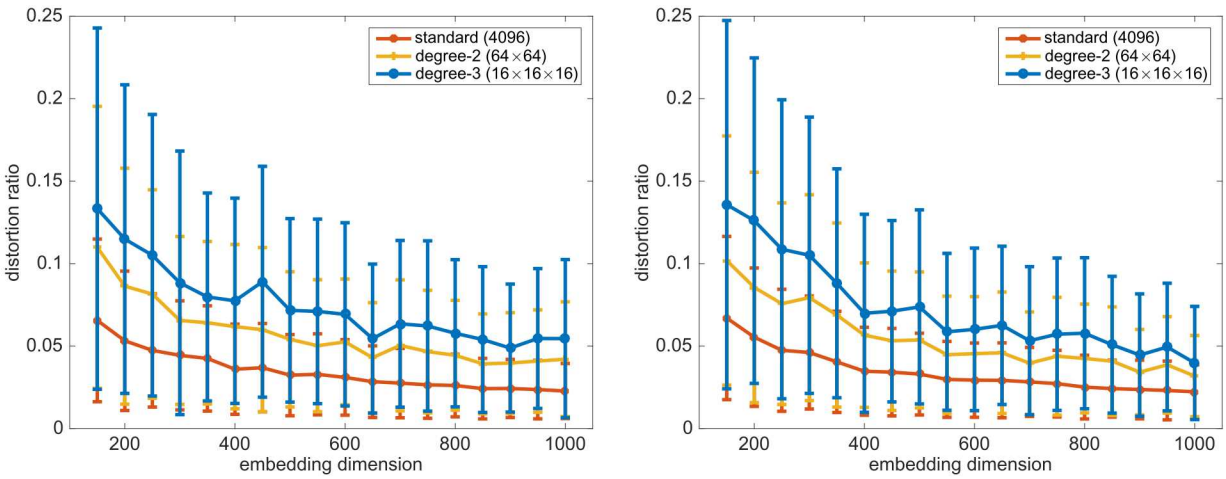
5.1. FJLT vs Kronecker FJLT. FJLT and KFJLT differ in the mixing operation. We show the numerical result Figure 2 comparing the embedding performance of standard FJLT, degree-2 and degree-3 KFJLTs on a set of randomly constructed Kronecker vectors. The numerical observation suggests that KFJLTs take slightly more rows to achieve the same quality of embeddings and lose some stability compared to standard FJLT, which is consistent with the theory.

5.2. Kronecker-structured vs general vectors. It is clearly of interest to study the general case of KFJLT embedding arbitrary Euclidean vectors since it is needed for the theoretical analysis of CPRAND-MIX algorithm, though KFJLT is designed to accelerate dimension-reduction for tall Kronecker-structured vectors. One might also wonder if the main embedding results can be improved if we just restrict to Kronecker vectors, but the experiments Figure 3 suggests that, the Kronecker-structured vectors result in worst-case embedding compared to general random vectors.

To understand how the Kronecker structure contributes to the gap, we go back to the technical proof. From the concentration inequality in Theorem 3.4, the probability bound in (3.3) is determined by

$$2\exp(-\frac{1}{88} \cdot \frac{\varepsilon}{\delta \cdot s^{d-2}}) + \frac{C_d}{4} \cdot \boldsymbol{\beta}_1$$

recalling $\boldsymbol{\beta}_1 = 4\exp(-s/128)$ is the probability bound for $|m_2(i,j)|$ concentrating in the scale 2δ for $i, j \in [n_2]$. Given a certain tolerance ε, it is more unlikely to control the overall distortion exceeding ε with a bigger $\boldsymbol{\beta}_1$.



(a) uniform $[0, 1]$ distributed vectors

(b) normally distributed vectors

Figure 2. *Comparing the embedding performance between the standard FJLT and the KFJLTs of degree* 2 *and* 3. *Each dot with an error bar represents the average distortion ratio and standard deviation based on* 1000 *trials for a given embedding dimension. In each trial, we generate the same subsampled FFT but different random sign-flipping operations for three constructions and test them on the same vector. The vectors to be embedded are* $(\mathbb{R}^4)^{\otimes 6}$ *Kronecker vectors, hence simultaneously degree-*2 *and degree-*3 *Kronecker vectors, and they consist of respectively uniform* $[0, 1]$ *and normally distributed elements in each component vector.*

When $i \neq j$, by a general version of Hanson-Wright inequality [35], $\boldsymbol{\beta}_1$ increases if $m_2(i, j)$ tends to concentrate around a greater expectation.

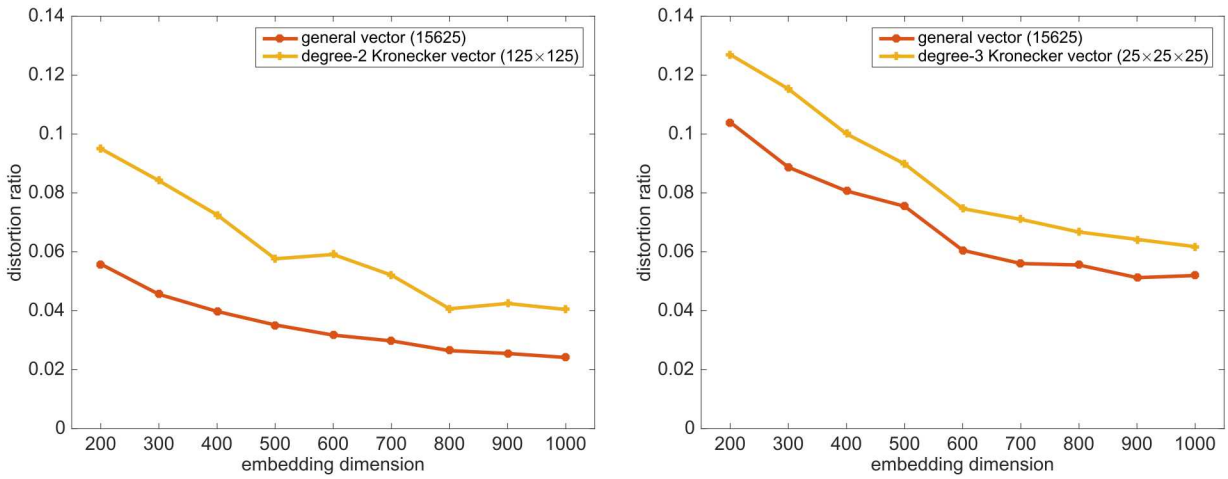
$$\mathbb{P}(|m_2(i,j) - \mathbb{E}\,(m_2(i,j))\,| > t) \leq 2\exp\ [-c \cdot \min(\frac{t^2}{\|\mathbf{M}_{i,j}\|_F^2}, \frac{t}{\|\mathbf{M}_{i,j}\|_2})]$$

Moreover,

$$\mathbb{E}\,(m_2(i,j)) = \sum_{i_1=1}^{n_1} x_i(i_1) \cdot x_j(i_1) \cdot \boldsymbol{\Psi}_i(i_1)^\top \boldsymbol{\Psi}_j(i_1),$$

the correlation between entries in blocks $\mathbf{x}_i$ and $\mathbf{x}_j$ can make a difference in the estimation of $\mathbb{E}\,(m_2(i,j))$. Following the *rearrangement inequality*, the expectation tends to reach its highest value among all the choices of pairwise arrangements when $x_j(i_1)$ is in the same position as $x_i(i_1)$ after reordering according to their decreasing arrangements. Vectors with Kronecker structure happen to be in this particular situation, thus achieving larger distortion in general, compared to general vectors.

5.3. Sampling strategy in KFJLT. In constructing the KFJLT, it might seem less natural to first construct the Kronecker product $\bigotimes_{k=d}^{1} \mathcal{F}_{n_k} \mathbf{D}_{n_k}$ and then subsample rows uniformly, as we propose, compared to first uniformly subsampling each $\mathcal{F}_{n_k}$ and then taking the Kronecker product of the resulting subsampled matrices. On the one hand, the sampling operation does not affect the computational savings for KFJLT, hence there is no major difference in the



(a) Degree-2 (125×125) KFJLT (b) Degree-3 ($25 \times 25 \times 25$) KFJLT

Figure 3. *Comparing the embedding performance of the KFJLT embedding on general and Kronecker Euclidean vectors. Each dot represents the average distortion ratio based on* 1000 *trials for a given embedding dimension. In each trial, we generate a general vector as well as a Kronecker vector and embed each of them using the same KFJLT. Each vector consists of normally distributed elements, either in full or each component vector.*

computational cost between two sampling methods. However, uniformly subsampling in the final step as we do does lead to a better JL embedding.

Indeed, consider instead sampling components $\mathbf{S}_1 \in \mathbb{R}^{m_1 \times n_1}, \mathbf{S}_2 \in \mathbb{R}^{m_2 \times n_2}, \cdots, \mathbf{S}_d \in \mathbb{R}^{m_d \times n_d}$, and forming the alternative embedding

$$\mathbf{\Phi x} = \left(\bigotimes_{k=d}^{1} \mathbf{S}_k \mathcal{F}_{n_k} \mathbf{D}_{n_k} \right) \left(\bigotimes_{k=d}^{1} \mathbf{x}_k \right) = \bigotimes_{k=d}^{1} \underbrace{(\mathbf{S}_k \mathcal{F}_{n_k} \mathbf{D}_{n_k})}_{\text{standard FJLT: } \mathbf{\Phi}_k} \mathbf{x}_k \tag{5.1}$$

We have the distortion estimation:

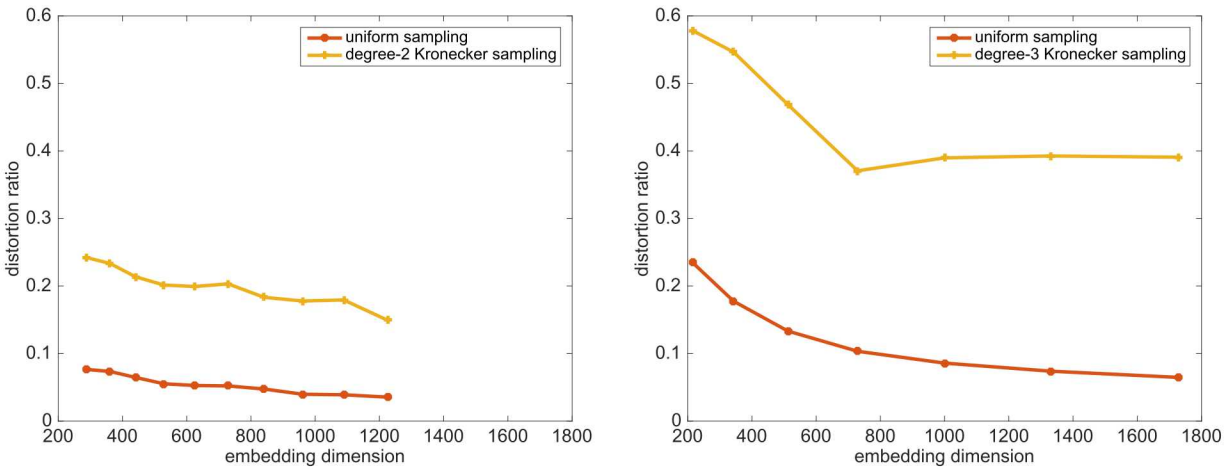
$$\|\mathbf{\Phi x}\|_2^2 - \|\mathbf{x}\|_2^2 = O\left(\max_{k \in [d]} \left| \|\mathbf{\Phi}_k \mathbf{x}_k\|_2^2 - \|\mathbf{x}_k\|_2^2 \right| \right)$$

To achieve a $(1 \pm \varepsilon)$ approximation, each m_k must be of the scale ε^{-2} based on (1.3). Hence the total embedding dimension $m = \prod_{k=1}^{d} m_k$ must be at least of the order ε^{-2d}, which is significantly worse than the scaling we obtain with uniform sampling, ε^{-2}.

We corroborate this calculation empirically below, comparing the distortions resulting from our KFJLT with those resulting from a Kronecker-factored sampling strategy as in (5.1).

Acknowledgments. The authors would like to thank Bosu Choi, Joe Neeman, Eric Price, Per-Gunnar Martinsson, David Woodruff and Gordan Žitković for valuable help and discussions.

REFERENCES



(a) Degree-2 (125×125) KFJLT (b) Degree-3 ($25 \times 25 \times 25$) KFJLT

Figure 4. *Comparing the embedding performance between the uniform and the Kronecker sampling strategies. Each dot represents the average distortion ratio based on* 1000 *trials for a given embedding dimension. In each trial, we generate the same sign-flipped FFT for each Kronecker component but different sampling instructions on the same embedding dimension for two constructions. We test them on the same vector. The embedded objects are respectively degree-*2 *and degree-*3 *Kronecker vectors in the two plots. They consist of normally distributed elements in each component vector.*

[1] N. Ailon and B. Chazelle, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*, in STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing, ACM, New York, 2006, pp. 557–563, https://doi.org/10.1145/1132516.1132597.

[2] N. Ailon and B. Chazelle, *The fast Johnson-Lindenstrauss transform and approximate nearest neighbors*, SIAM J. Comput., 39 (2009), pp. 302–322, https://doi.org/10.1137/060673096.

[3] N. Ailon and E. Liberty, *An almost optimal unrestricted fast johnson-lindenstrauss transform*, ACM Trans. Algorithms, 9 (2013), pp. 21:1–21:12, https://doi.org/10.1145/2483699.2483701.

[4] H. Avron, H. Nguyen, and D. Woodruff, *Subspace embeddings for the polynomial kernel*, in Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2258–2266, http://papers.nips.cc/paper/5240-subspace-embeddings-for-the-polynomial-kernel.pdf.

[5] B. W. Bader, T. G. Kolda, et al., *Matlab tensor toolbox version 3.1*. Available online, June 2019, https://www.tensortoolbox.org.

[6] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx., 28 (2008), pp. 253–263, https://doi.org/10.1007/s00365-007-9003-x.

[7] C. Battaglino, G. Ballard, and T. G. Kolda, *A practical randomized CP tensor decomposition*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 876–901, https://doi.org/10.1137/17M1112303.

[8] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities using the entropy method*, Ann. Probab., 31 (2003), pp. 1583–1614, https://doi.org/10.1214/aop/1055425791.

[9] J. Bourgain, *An improved estimate in the restricted isometry problem*, in Geometric aspects of functional analysis, vol. 2116 of Lecture Notes in Math., Springer, Cham, 2014, pp. 65–70, https://doi.org/10.1007/978-3-319-09477-9_5.

[10] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509, https://doi.org/10.1109/TIT.2005.862083.

[11] E. J. Candes and T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425, https://doi.org/10.1109/TIT.2006.885507.

[12] M. Charikar, K. C. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, Theor. Comput. Sci., 312 (2004), pp. 3–15, https://doi.org/10.1016/S0304-3975(03)00400-6, https://doi.org/10.1016/S0304-3975(03)00400-6.

[13] M. Cheraghchi, V. Guruswami, and A. Velingker, *Restricted isometry of Fourier matrices and list decodability of random linear codes*, SIAM J. Comput., 42 (2013), pp. 1888–1914, https://doi.org/10.1137/120896773.

[14] K. L. Clarkson and D. P. Woodruff, *Low rank approximation and regression in input sparsity time*, in Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC '13, New York, NY, USA, 2013, ACM, pp. 81–90, https://doi.org/10.1145/2488608.2488620.

[15] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms, 22 (2003), pp. 60–65, https://doi.org/10.1002/rsa.10073.

[16] H. Diao, Z. Song, W. Sun, and D. P. Woodruff, *Sketching for kronecker product regression and p-splines*, in International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018, pp. 1299–1308, http://proceedings.mlr.press/v84/diao18a.html.

[17] M. F. Duarte and R. G. Baraniuk, *Kronecker compressive sensing*, IEEE Transactions on Image Processing, 21 (2012), pp. 494–504, https://doi.org/10.1109/TIP.2011.2165289.

[18] A. Dutt and V. Rokhlin, *Fast fourier transforms for nonequispaced data*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1368–1393, https://doi.org/10.1137/0914081.

[19] A. Fijany and C. P. Williams, *Quantum wavelet transforms: Fast algorithms and complete circuits*, in Selected Papers from the First NASA International Conference on Quantum Computing and Quantum Communications, London, UK, 1998, Springer-Verlag, pp. 10–33, https://doi.org/10.1007/3-540-49208-9_2.

[20] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, New York, 2013, https://doi.org/10.1007/978-0-8176-4948-7.

[21] G. G. Lorentz, M. V. Golitschek, and Y. Makovoz, *Constructive Approximation (Advanced Problems)*.

[22] D. L. Hanson and F. T. Wright, *A bound on tail probabilities for quadratic forms in independent random variables*, Ann. Math. Statist., 42 (1971), pp. 1079–1083, https://doi.org/10.1214/aoms/1177693335.

[23] I. Haviv and O. Regev, *The restricted isometry property of subsampled fourier matrices*, in Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 288–297, https://doi.org/10.1137/1.9781611974331.ch22.

[24] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. American Statist. Assoc., 58 (1963), pp. 13 – 30, https://doi.org/10.2307/2282952.

[25] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Conference in modern analysis and probability (New Haven, Conn., 1982), vol. 26 of Contemp. Math., Amer. Math. Soc., Providence, RI, 1984, pp. 189–206, https://doi.org/10.1090/conm/026/737400.

[26] F. Krahmer, S. Mendelson, and H. Rauhut, *Suprema of chaos processes and the restricted isometry property*, Comm. Pure Appl. Math., 67 (2014), pp. 1877–1904, https://doi.org/10.1002/cpa.21504.

[27] F. Krahmer and R. Ward, *New and improved johnson-lindenstrauss embeddings via the restricted isometry property*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 1269–1281, https://doi.org/10.1137/100810447.

[28] K. G. Larsen and J. Nelson, *Optimality of the johnson-lindenstrauss lemma*, in 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), Oct 2017, pp. 633–638, https://doi.org/10.1109/FOCS.2017.64.

[29] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, *Kronecker graphs: An approach to modeling networks*, J. Mach. Learn. Res., 11 (2010), pp. 985–1042, https://doi.org/10.1145/1756006.1756039.

[30] J. Martens and R. Grosse, *Optimizing neural networks with kronecker-factored approximate curvature*, in Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, pp. 2408–2417, http://dl.acm.org/citation.cfm?id=3045118.3045374.

[31] X. Meng and M. W. Mahoney, *Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression*, in Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC '13, New York, NY, USA, 2013, ACM, pp. 91–100, https://doi.org/10.1145/2488608.2488621.

[32] N. Pham and R. 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, New York, NY, USA, 2013, ACM, pp. 239–247, https://doi.org/10.1145/2487575.2487591.

[33] H. Rauhut, *Compressive sensing and structured random matrices*, in Theoretical foundations and numerical methods for sparse recovery, vol. 9 of Radon Ser. Comput. Appl. Math., Walter de Gruyter, Berlin, 2010, pp. 1–92, https://doi.org/10.1515/9783110226157.1.

[34] P. A. Regalia and S. K. Mitra, *Kronecker products, unitary matrices and signal processing applications*, SIAM Review, 31 (1989), https://doi.org/10.1137/1031127.

[35] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math., 61 (2008), pp. 1025–1045, https://doi.org/10.1002/cpa.20227.

[36] Y. Sun, Y. Guo, J. A. Tropp, and M. Udell, *Tensor random projection for low memory dimension reduction*, in NeurIPS Workshop on Relational Representation Learning, 2018, https://r2learning.github.io/assets/papers/CameraReadySubmission%2041.pdf.

[37] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100, https://doi.org/10.1016/S0377-0427(00)00393-9. Numerical analysis 2000, Vol. III. Linear algebra.

[38] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. iv+157, https://doi.org/10.1561/0400000060.

Appendix A. Proof of Proposition 4.6.

Proof. Due to the $(2s, \delta)$-RIP property of $\mathbf{\Psi}$, apply the result from Lemma 4.2, for any row and column index sets $I, J \subset [n]$, $\|\mathbf{\Psi}_L(I)^\top \mathbf{\Psi}_R(J)\| \leq \delta$, if $|I| \leq s, |J| \leq s$.

$$
\begin{aligned}
\|\mathbf{C}_{\mathbf{x},\mathbf{y}}\| &= \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} |\langle \mathbf{u}, \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{v}\rangle| \\
&\leq \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \sum_{\substack{p,q=2 \\ p\neq q}}^{r} |\langle \mathbf{u}(I_p), \mathbf{C}_{\mathbf{x},\mathbf{y}}(I_p, J_q)\mathbf{v}(J_q)\rangle| \\
&\leq \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \sum_{\substack{p,q=2 \\ p\neq q}}^{r} \|\mathbf{u}(I_p)\|_2 \cdot \|\mathbf{v}(J_q)\|_2 \cdot \|\mathbf{D}_{\mathbf{x}(I_p)}\mathbf{\Psi}_L(I_p)^\top \mathbf{\Psi}_R(J_q)\mathbf{D}_{\mathbf{y}(J_q)}\| \\
&\leq \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \sum_{\substack{p,q=2 \\ p\neq q}}^{r} \|\mathbf{u}(I_p)\|_2 \cdot \|\mathbf{v}(J_q)\|_2 \cdot \|\mathbf{x}(I_p)\|_\infty \cdot \|\mathbf{y}(J_q)\|_\infty \cdot \delta \\
&\leq \delta \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \sum_{\substack{p,q=2 \\ p\neq q}}^{r} \|\mathbf{u}(I_p)\|_2 \cdot \|\mathbf{v}(J_q)\|_2 \cdot \frac{1}{\sqrt{s}} \cdot \|\mathbf{x}(I_{p-1})\|_2 \cdot \frac{1}{\sqrt{s}} \cdot \|\mathbf{y}(J_{q-1})\|_2 \\
&\leq \frac{\delta}{s} \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \sum_{\substack{p,q=2 \\ p\neq q}}^{r} \left(\frac{1}{2}\|\mathbf{u}(I_p)\|_2^2 + \frac{1}{2}\frac{\|\mathbf{x}(I_{p-1})\|_2^2}{\|\mathbf{x}\|_2^2}\right) \cdot \left(\frac{1}{2}\|\mathbf{v}(J_q)\|_2^2 + \frac{1}{2}\frac{\|\mathbf{y}(J_{q-1})\|_2^2}{\|\mathbf{y}\|_2^2}\right) \\
&\leq \frac{\delta}{s} \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2.
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbf{C}_{\mathbf{x},\mathbf{y}}\|_F^2 &= \sum_{\substack{i \not\sim j \\ i\in I_1^c, j\in J_1^c}}^{n} \Big(x(i)\mathbf{\Psi}_L(i)^\top \mathbf{\Psi}_R(j) y(j)\Big)^2 \\
&= \sum_{q=2}^{r} \sum_{\substack{i\in I_1^c \\ i\notin J_q}}^{n} x(i)^2 \|\mathbf{D}_{\mathbf{y}(J_q)}\mathbf{\Psi}_R(J_q)^\top \mathbf{\Psi}_L(i)\|^2 \\
&\leq \sum_{q=2}^{r} \sum_{\substack{i\in I_1^c \\ i\notin J_q}}^{n} x(i)^2 \cdot \|\mathbf{y}(J_q)\|_\infty^2 \cdot \|\mathbf{\Psi}_R(J_q)^\top \mathbf{\Psi}_L(i)\|^2 \\
&\leq \sum_{q=2}^{r} \frac{\delta^2}{s} \cdot \|\mathbf{y}(J_{q-1})\|_2^2 \cdot \sum_{i=1}^{n} x(i)^2 \\
&\leq \frac{\delta^2}{s} \cdot \|\mathbf{x}\|_2^2 \cdot \|\mathbf{y}\|_2^2.
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbf{v}_{\mathbf{x},\mathbf{y}}\|_2 &\leq \sup_{\|\mathbf{u}\|_2=1} \sum_{p=2}^{r} \langle \mathbf{u}(I_p), \mathbf{D}_{\mathbf{x}(I_p)} \mathbf{\Psi}_L(I_p)^\top \mathbf{\Psi}_R(J_1) \mathbf{D}_{\mathbf{b}} \mathbf{y}(J_1) \rangle \\
&\leq \sup_{\|\mathbf{u}\|_2=1} \sum_{p=2}^{r} \|\mathbf{u}(I_p)\|_2 \cdot \|\mathbf{x}(I_p)\|_\infty \cdot \|\mathbf{b}\|_\infty \cdot \|\mathbf{\Psi}_L(I_p)^\top \mathbf{\Psi}_R(J_1)\| \cdot \|\mathbf{y}(J_1)\|_2 \\
&\leq \sup_{\|\mathbf{u}\|_2=1} \sum_{p=2}^{r} \|\mathbf{u}(I_p)\|_2 \cdot \frac{1}{\sqrt{s}} \cdot \|\mathbf{x}(I_{p-1})\|_2 \cdot \delta \cdot \|\mathbf{y}\|_2 \\
&\leq \frac{\delta}{\sqrt{s}} \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \sup_{\|\mathbf{u}\|_2=1} \sum_{p=2}^{r} \left(\frac{1}{2} \|\mathbf{u}(I_p)\|_2^2 + \frac{1}{2} \frac{\|\mathbf{x}(I_{p-1})\|_2^2}{\|\mathbf{x}\|_2^2} \right) \\
&\leq \frac{\delta}{\sqrt{s}} \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2
\end{aligned}
$$

$$
\begin{aligned}
|\mathbf{W}_{\mathbf{x},\mathbf{y}}| &\leq \sum_{p=1}^{r} \left| \mathbf{d}(I_p)^\top \mathbf{D}_{\mathbf{x}(I_p)} \mathbf{\Psi}_L(I_p)^\top \mathbf{\Psi}_R(J_p) \mathbf{D}_{\mathbf{y}(J_p)} \mathbf{d}(J_p) \right| \\
&\leq \sum_{p=1}^{r} \|\mathbf{x}(I_p)\|_2 \cdot \|\mathbf{y}(J_p)\|_2 \cdot \|\mathbf{\Psi}_L(I_p)^\top \mathbf{\Psi}_R(J_p)\|, \\
&\leq \delta \sum_{p=1}^{r} \|\mathbf{x}(I_p)\|_2 \cdot \|\mathbf{y}(J_p)\|_2 \\
&\leq \delta \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \sum_{p=1}^{r} \frac{1}{2} \left(\frac{\|\mathbf{x}(I_p)\|_2^2}{\|\mathbf{x}\|_2^2} + \frac{\|\mathbf{y}(J_p)\|_2^2}{\|\mathbf{y}\|_2^2} \right) \\
&= \delta \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2.
\end{aligned}
$$

■

Appendix B. Fitting CP model with alternating randomized least squares. In this section, we give supplemental material on the CPRAND-MIX algorithm and show that the application of Kronecker FJLT to the alternating least squares problem greatly reduces the workload of CP tensor decomposition.

The Khatri-Rao product, also called the column-wise Kronecker product denoted by $\odot$, is defined as: given matrices $\mathbf{U} \in \mathbb{R}^{I \times J}$ and $\mathbf{V} \in \mathbb{R}^{K \times J}$,

$$
\mathbf{U} \odot \mathbf{V} = \begin{bmatrix} \mathbf{u}_1 \otimes \mathbf{v}_1 & \mathbf{u}_2 \otimes \mathbf{v}_2 & \cdots & \mathbf{u}_J \otimes \mathbf{v}_J \end{bmatrix} \in \mathbb{R}^{IK \times J}. \tag{B.1}
$$

The Khatri-Rao product also satisfies the distributive property

$$
\mathbf{UV} \odot \mathbf{TW} = (\mathbf{U} \otimes \mathbf{T})(\mathbf{V} \odot \mathbf{W}). \tag{B.2}
$$

B.1. Problem set-up. Let $\mathcal{X}$ be a d-way tensor of size $n_1 \times n_2 \times \cdots \times n_d$, and $\mathcal{M}$ be a low-rank approximation of $\mathcal{X}$ such that $\text{rank}(\mathcal{M}) \leq r$. $\mathcal{M}$ is defined by d factor matrices, i.e.

$\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$ via

$$\mathcal{M} = \sum_{j=1}^{r} \mathbf{A}_1(:,j) \circ \mathbf{A}_2(:,j) \circ \cdots \circ \mathbf{A}_d(:,j). \tag{B.3}$$

The goal of fitting tensor CP model is to find the factor matrices that minimize the nonlinear least squares objective:

$$\|\mathcal{X} - \mathcal{M}\|^2 = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} (x(i_1, i_2, \cdots, i_d) - m(i_1, i_2, \cdots, i_d))^2 \tag{B.4}$$

subject to $\mathcal{M}$ being low rank as (B.3).The tensor $\mathcal{X}$ has $\prod_{k=1}^{d} n_k$ parameters whereas $\mathcal{M}$ has only $r \sum_{k=1}^{d} n_k$ parameters.

For ease of the notation, we define $N = \prod_{k=1}^{d} n_k$ and $N_k = N/n_k$.

The mode-k unfolding of $\mathcal{X}$ recognizes the elements of the tensor into a matrix $\mathbf{X}_{(k)}$ of size $n_k \times N_k$. The mode-k unfolding of $\mathcal{M}$ has a special structure:

$$\mathcal{M}_{(k)} = \mathbf{A}_k \underbrace{(\mathbf{A}_d \odot \cdots \odot \mathbf{A}_{k+1} \odot \mathbf{A}_{k-1} \cdots \odot \mathbf{A}_1)^\top}_{\mathbf{Z}_k^\top}. \tag{B.5}$$

The idea behind *alternating least squares* (ALS) for CP is solving for one factor matrix $\mathbf{A}_k$ at a time, repeating the cycle until the method converges. This takes advantage of the fact that we can rewrite the minimization problem using (B.5) as

$$\min_{\mathbf{A}_k} \|\mathbf{Z}_k \mathbf{A}_k^\top - \mathbf{X}_{(k)}^\top\|_F, \tag{B.6}$$

which is a linear least square problem with a closed form solution. The cost of solving the least square problem is $O(rN)$ due to the particular structure of $\mathbf{Z}_k$. But we need to solve d such problems per outer loop and run tens or hundreds of outer loops to solve a typical CP-ALS problem. Hence, reducing the cost of (B.6) is of interest.

B.2. Randomized least squares. Since we expect that the number of rows N_k is much greater than the number of columns r in $\mathbf{Z}_k$, (B.6) can benefit from randomized sketching methods. Instead of solving the full least square, we can instead solve a reduced problem by a sketch matrix $\mathbf{\Phi} \in \mathbb{R}^{\mathbf{s} \times N_k}$:

$$\min_{\mathbf{A}_k} \|\mathbf{\Phi} \mathbf{Z}_k \mathbf{A}_k^\top - \mathbf{\Phi} \mathbf{X}_{(k)}^\top\|. \tag{B.7}$$

For $\mathbf{\Phi}$ being a FJLT, the dominant cost are applying the FFT to $\mathbf{Z}_k$ and $\mathbf{X}_{(k)}$ and solving the least square: $O((r + n_k) N_k \log N_k + r^2 \mathbf{s})$.

We then change the sketching form of $\mathbf{\Phi}$ to be a Kronecker FJLT:

$$\mathbf{\Phi} = \mathbf{S} \left(\bigotimes_{\substack{\ell=d \\ \ell \neq k}}^{1} \mathcal{F}_{n_\ell} \mathbf{D}_{n_\ell} \right), \tag{B.8}$$

as we can compute (B.6) more efficiently.

First consider the multiplication with $\mathbf{X}_{(k)}^\top$. We pay an *one-time* upfront cost to reduce the cost per iteration. The corresponding computation is to mix the original tensor:

$$\hat{\mathcal{X}} = \mathcal{X} \times_1 \mathcal{F}_{n_1}\mathbf{D}_{n_1} \cdots \times_d \mathcal{F}_{n_d}\mathbf{D}_{n_d}. \tag{B.9}$$

The total cost is $N \log N$.

We observe that

$$\mathbf{\Phi}\mathbf{X}_{(k)}^\top = \left(\mathbf{S}\hat{\mathbf{X}}_{(k)}^\top\right)\mathcal{F}_{n_k}^*\mathbf{D}_{n_k}. \tag{B.10}$$

The asterisk $*$ denotes the conjugate transpose. This equation shows that we just need to sample and then apply the inverse FFT and diagonal. The work per iteration is $O(\mathbf{s}n_k \log n_k)$.

Next consider $\mathbf{\Phi}\mathbf{Z}_k$. We finish the mixing for $\mathbf{A}_k$: $\hat{\mathbf{A}}_k = \mathcal{F}_{n_k}\mathbf{D}_{n_k}\mathbf{A}_k$, which costs $O(rn_k \log n_k)$, before sketching the least square in mode k. Then the cost of computing

$$\mathbf{\Phi}\mathbf{Z}_k = \mathbf{S}\left(\bigodot_{\substack{\ell=d\\ \ell\neq k}}^{1} \mathcal{F}_{n_\ell}\mathbf{D}_{n_\ell}\mathbf{A}_\ell\right) = \mathbf{S}\left(\bigodot_{\substack{\ell=d\\ \ell\neq k}}^{1} \hat{\mathbf{A}}_\ell\right) \tag{B.11}$$

is just the cost of sampling the Khatri-Rao product: $r\mathbf{s}$.

To conclude the comparison of the cost in Table 2:

Table 2
Cost per inner iteration

Regular CP-ALS	FJLT - sketched	Kronecker FJLT - sketched
$O(rN)$	$O((r+n_k)N_k \log(N_k) + r^2\mathbf{s})$	$O((r+\mathbf{s})n_k \log n_k + r^2\mathbf{s})$

It is natural to choose Kronecker FJLT as the sketch strategy for solving CP alternating least squares, as it helps reduce the cost of the inner iteration greatly to the order of $O(n_k \log n_k)$ compared to the original cost: $O(N)$. This idea has been developed into a randomized algorithm: CPRAND-MIX. We refer the readers to [7] for the completed algorithm.