High Probability Tensor Sketch

Thomas Dybdahl Ahle

IT University of Copenhagen

Jakob Bæk Tejs Knudsen

Copenhagen University

Abstract

We construct a structured Johnson Lindenstrauss transformation that can be applied to simple tensors on the form $x = x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ in time nearly cd. That is, exponentially faster than writing out the Kronecker product and then mapping down.

These matrices, M, which preserves the norm of any $x \in \mathbb{R}^{d^c}$, such that $|\|Mx\|_2 - \|x\||_2 \le \epsilon$ with probability $1-\delta$, can be taken to have just $\tilde{O}(c^2\epsilon^{-2}(\log 1/\delta)^3)$ rows. This is within $c^2(\log 1/\delta)^2$ of optimal for any JL matrix [16], and improves upon earlier 'Tensor Sketch' constructions by Pagh and Pham [24, 25], which used $\tilde{O}(3^c\epsilon^{-2}\delta^{-1})$ rows, by an exponential amount in both c and δ^{-1}.

It was shown by Avron, Nguyen and Woodruff in [5] that Tensor Sketch is a subspace embedding. This has a large number of applications [28], such as guaranteeing the correctness of kernel-linear regression performed directly on the reduced vectors. We show that our construction is a subspace embedding too, improving again upon the exponential dependency on c and δ^{-1}, enabling sketching of much higher order polynomial kernels, such as Taylor approximations to the ubiquitous Gaussian radial basis function.

Technically, we construct our matrix M such that $M(x \otimes y) = Tx \circ T'y$ where $\circ$ is the Hadamard (element-wise) product and T and T' support fast matrix-vector multiplication ala [1]. To analyze the behavior of Mx on non-simple x, we show a higher order version of Khintchine's inequality [14], related to the higher order Gaussian chaos analysis by Latała [19, 17]. Finally we show that such sketches can be combined recursively, in a way that doesn't increase the dependency on c by much.

2012 ACM Subject Classification Theory of computation → Sketching and sampling

Keywords and phrases Tensors, Sketching, Johnson Lindenstrauss

Digital Object Identifier 10.4230/LIPIcs...

Acknowledgements Anders Aamand

1 Introduction

The polynomial method has recently found many great applications in algorithm design, such as finding orthogonal vectors [2] and gap amplification in nearest neighbour [27]. Consider the polynomial $P(x, y) = \sum_{i<j<k}(x_i + y_i - x_iy_i)(x_j + y_j - x_jy_j)(x_k + y_k - x_ky_k)$ which counts the number of triangles in the union between two graphs, $x, y \in \{0, 1\}^{\binom{n}{2}}$, expressed as binary vectors over the edges. Splitting P into monomials, one may construct functions $f, g : \{0, 1\}^{\binom{n}{2}} \to \mathbb{R}^m$ such that $P(x, y) = \langle f(x), g(y)\rangle$, and use these embeddings to solve the 'most triangles in union' problem on a database of graphs, using an off the shelf nearest neighbours algorithm. (See our Applications for more on this.)

Other examples are kernel functions in statistics, such as $P(x, y) = \exp(-\|x - y\|_2^2) = \sum_{k\ge0}(-1)^k\|x - y\|^{2k}/k!$, the Gaussian Radial Basis Function. The celebrated 'kernel trick' has been used in linear methods such as kernel PCA kernel nearest neighbour or kernel regression to allow detection of nonlinear dependencies between data without explicitly constructing feature vectors in high dimensional spaces. However the 'trick' requires the computation of the inner product between all pairs of points, while explicit embeddings scale linearly in the number of points, and have thus recently experienced a comeback.

While these polynomial expansions often produce prohibitively large vectors, they can often be reduced by some means, such as the Johnson-Lindenstrauss transform [13]. This in turn creates a strange phenomenon where we first blow up the dimension only to later squash it back down. It is tempting to look for shortcut to go straight to the final dimension.

This was the idea of Pagh and Pham [24, 25] with Tensor Sketch. The observation was that $P(x, y) = \langle x, y\rangle^2 = \langle x \otimes x, y \otimes y\rangle$, where $x \otimes x$ is the tensor product (Kronecker) of x with itself. They further observed that if C and C' are independent Count Sketch matrices, then $\langle Cx * C'x, Cy * C'y\rangle \approx \langle x, y\rangle^2$ while the dimension of $Cx * C'x$ is much smaller than of $x \otimes x$. Since the convolution $(\cdot * \cdot)$ can be computed in near linear time using the fast Fourier transformation, they could sketch $x \otimes x$ in basically the time required to sketch x. For a higher order polynomial kernel, replacing $f(x) = x^{\otimes c}$ with $f(x) = C^{(1)}x * \cdots * C^{(c)}x$ thus takes the sketching time from d^c to cd. A huge improvement!

In this paper we improve on the main shortcomings of Tensor Sketch: To preserve the norm of vectors up to $1 \pm \epsilon$ with probability $1 - \delta$, it requires embedding into dimension roughly $3^c\epsilon^{-2}\delta^{-1}$. The exponential dependency on c greatly limits the degree of polynomials that can be embedded, and the linear dependency on δ^{-1} means we can't use a standard union bound trick to get e.g. a near neighbour preserving embedding [11], as could be achieved with the Johnson Lindenstrauss transform, which embeds into only $\epsilon^{-2}\log 1/\delta$ dimensions. We overcome both of these obstacles, by analyzing a scheme, that with the same embedding time, requires only $c^2\epsilon^{-2}(\log 1/\delta)^3$ dimensions.

A hugely important idea was introduced by Avron et al. [5]: They proved that a Tensor Sketch with sufficiently many rows is a Subspace embedding. This allowed many application that previously were only applied heuristically, such as solving a regression problem $\arg\min_x \|Ax - y\|_2$ directly in the reduced space while guaranteeing correct results. Using the subspace embeddings, they obtained the fastest known algorithms for computing an computing an approximate kernel PCA and many other problems.

However, the weaknesses of Tensor Sketch remained: The exponential dependency on c meant that the method could only be applied with relatively low degree polynomials. In this paper we show that our High Probability Tensor Sketch is also a subspace embedding, solving this major roadblock. We also show a second version of our sketch, which improves upon [5] by allowing an embedding dimension linear in the subspace dimension, rather than quadratic. In many uses of the subspace method the embedding dimension becomes larger than the number of points, which means we can get a quadratic improvement on these applications.

Our approach is to analyze fast family of Johnson Lindenstrauss matrices M with the further property that $M(x \otimes y) = M'x \circ M''y$ where $\circ$ is the Hadamard (or element-wise) vector product. We also analyze the case where M' and M'' are fully independent Gaussian matrices, and show that we are within a single factor $\log 1/\delta$ in embedding dimension while supporting much faster matrix-vector multiplication. The direct application of this method, $M^{(1)}x^{(1)} \circ \cdots \circ M^{(c)}$, would result in an exponential dependency on c, but by instead combining vectors as $M^{(1)}x^{(1)} \circ M'^{(1)}(M^{(2)}x^{(2)} \circ M'^{(2)}(\ldots$ we show that this dependency can be reduced to c^2. See also the Technical Overview below.

1.1 Overview

Our main contribution is to answer the questions "Does Tensor Sketch work with high probability?" and "Does there exist subspace embeddings for higher order polynomial embeddings?" For both of those questions, the answer is yes!

▶ **Theorem 1** (Construction A). *There is a distribution $\mathcal{M}$ over matrices $M \in \mathbb{R}^{m\times d^c}$ where*

1. $|\|Mx\|_2 - \|x\|_2| \le \epsilon$ *with probability* $\ge 1-\delta$ *for any* $x \in \mathbb{R}^{d^c}$.
2. M *can be applied to tensors* $x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ *in time* $O(c\,(d\log d + m\log m))$.
3. m *can be taken to be* $O(c^2\epsilon^{-2}(\log 1/\delta)(\log 1/\epsilon\delta)^2)$.
4. $\mathcal{M}$ *can compute fast approximate matrix multiplication:* $\Pr[|\|A^T M^T M B - A^T B\|_F| > \epsilon\|A\|_F\|B\|_F] < \delta$.
5. *There is an* $m = O(c^2\epsilon^{-2}\lambda^2(\log 1/\delta)(\log 1/\epsilon\delta)^2)$ *such that* $\mathcal{M}$ *is an* (ϵ,δ)*-subspace embedding. (See definition 11.)*

The result matches, up to a single factor $\log 1/\delta$ the embedding dimension needed for fully independent Gaussian matrices M and M', for $\|Mx \circ M'y\|_2$ to approximate $\|x \otimes y\|_2$. (See Appendix, Theorem 27.) However, suffering slightly in the embedding time, we can go all the way down to one:

▶ **Theorem 2** (Construction B). *There is a distribution* $\mathcal{M}$ *over matrices* $M \in \mathbb{R}^{m\times d^c}$ *where*

1. $|\|Mx\|_2 - \|x\|_2| \le \epsilon$ *with probability* $\ge 1-\delta$ *for any* $x \in \mathbb{R}^{d^c}$.
2. *Matrices* $M \sim \mathcal{M}$ *can be applied to tensors* $x^{(1)}\otimes\cdots\otimes x^{(c)} \in \mathbb{R}^{d^c}$ *in time* $O(c\,m\,\min(d,m))$.
3. m *can be taken to be* $O(c^2\epsilon^{-2}\log 1/\delta\,\log^2(c\epsilon^{-1}\log 1/\delta))$.
4. *There is an* $m = O(c^2\epsilon^{-2}(\lambda + \log 1/\delta)\log^2(c\epsilon^{-1}\lambda\log 1/\delta))$ *such that* $\mathcal{M}$ *is an* (ϵ,δ)*-subspace embedding. (See definition 11.)*

While the first theorem requires an intricate analysis of the combination of two Fast-JL matrices, the second one follows nearly directly from our general recursive construction theorem below:

▶ **Theorem 3.** *Let* $c > 0$ *be a positive integer, and* $Q^{(1)} \in \mathbb{R}^{m\times d}$ *and* $Q^{(i)} \in \mathbb{R}^{m\times md}$ *be independent random matrices for every* $i \in [c]\setminus\{1\}$*. Define* $M^{(k)} = Q^{(k)}(M^{(k-1)} \otimes I_d) \in \mathbb{R}^{m\times d^k}$ *for* $k \in [c]$*, where* $M^{(0)} = 1 \in \mathbb{R}$*. Let* $t > 0$ *be a positive integer, and let* $k_i \in [c]$ *for every* $i \in [t]$*. Then the matrix*

$$M = \bigoplus_{i\in[t]} M^{(k_i)} \in \mathbb{R}^{tm\times\Sigma_{i\in[t]}d^{k_i}}$$

has the following properties.

1. *Let* $\varepsilon \in (0,1)$ *and* $\delta > 0$*. If* $Q^{(i)}$ *has* $(\varepsilon/2c, \delta/c)$*-JL property for every* $i \in [c]$*, then* M *has* (ε,δ)*-JL property.*
2. *If* $Q^{(i)}x$ *can be evaluated in time* T*, where* $x \in \mathbb{R}^{md}$*, for every* $i \in [c]\setminus\{1\}$*, and* $Q^{(1)}y$ *can be evaluated in time* T'*, where* $y \in \mathbb{R}^d$*, then* $M(\bigoplus_{i\in[t]}\bigotimes_{j\in[k_i]} x^{(i,j)})$ *can be evaluated in time* $O(T't + T\sum_{i\in[t]} k_i)$*, where* $x^{(i,j)} \in \mathbb{R}^d$ *for every* $i \in [t], j \in [k_i]$*.*

Now the difference between Construction A and Construction B is simply which matrices $Q^{(1)},\ldots,Q^{(c))}$ that are used as basis for the construction.

Paper Structure

The paper is structured as follows: After the comparison to related work and preliminaries we give a Technical overview of the sketch. We find it is useful to have some established notation before this section.

The technical part is split in three: We first show Theorem 3. This gives a recursive construction, which can be applied to tensors using any of the shelf Johnson-Lindenstrauss matrix. Combined with the fastest analysis of Fast-JL [15] this gives our theorem 2.

We proceed to analyze a small change in the construction of Fast-JL matrices, which allow for very fast application to tensor products. Specifically we show that the random diagonal matrix can be replaced by the Kronecker product of two smaller diagonal matrices without losing the JL-property, if the number of rows is increased slightly. This gives our theorem 1.

Finally in the last section, we show some algorithmic applications of our constructions. For example how to use it to find the two graphs in a database whose union has the most triangles.

1.2 Related work

Work related to sketching of tensors and explicit kernel embeddings is found in fields ranging from pure mathematics to physics and machine learning. Hence we only try to compare ourselves with the four most common types we have found.

We focus particularly on the work on subspace embeddings [25, 5], since it is most directly comparable to ours. An extra entry in this category is [4], which is currently in review, and which we were made aware of while writing this paper. That work is in double blind review, but by the time of the final version of this paper, we should be able to cite it properly.

Subspace embeddings

For most applications [5], the subspace dimension, λ, will be much larger than the input dimension, d, but smaller than the implicit dimension d^c. Hence the size of the sketch, m, will also be assumed to satisfy $d << m << d^c$ for the purposes of stating the results. We will hide constant factors, and $\log 1/\epsilon$, $\log d$, $\log m$, $\log c$, $\log \lambda$ factors.

Note that we can always go from m down to $\approx \epsilon^{-2}(\lambda + \log 1/\delta)$ by applying a fast-JL transformation after embedding. This works because the product of two subspace embeddings is also a subspace embedding, and because fast-JL is a subspace embedding by the net-argument (see lemma 13). The embedding dimensions in the table should thus mainly be seen as a space dependency, rather than the actual embedding dimension for applications.

Reference	Embedding dimension, m	Embedding time	Note
[25, 5]	$\tilde{O}(3^c\, d\, \lambda^2\, \delta^{-1}\, \epsilon^{-2})$	$\tilde{O}(c\,(d+m))$	
Theorem 1	$\tilde{O}(c^2\, \lambda^2\, (\log 1/\delta)^3\, \epsilon^{-2})$	$\tilde{O}(c\,(d+m))$	
Theorem 2	$\tilde{O}(c^2\, (\lambda + \log 1/\delta)\, \epsilon^{-2})$	$\tilde{O}(c\, d\, m)$	
[4], Theorem 1	$\tilde{O}(c\, \lambda^2\, \delta^{-1}\, \epsilon^{-2})$	$\tilde{O}(c\,(d+m))$	Independent work.
[4], Theorem 2	$\tilde{O}(c^6\, \lambda\, (\log 1/\delta)^5\, \epsilon^{-2})$	$\tilde{O}(c\,(d+m))$	Independent work

Some of the results, in particular [25, 5] and [4] Theorem 1 can be applied faster when the input is sparse. Our results, as well as [4], Theorem 2 can similarly be optimized for sparse inputs, by preprocessing vectors with an implementation of Sparse JL [7].

In comparison to the previous result [25, 5] we are clearly better with an exponential improvement in c as well as δ.

Compared to the new work of [4], all four bounds have some region of superiority. Their first bound of has the best dependency on c, but has an exponential dependency on $\log 1/\delta$. Their second bound has an only linear dependency on $d + \lambda$, but has large polynomial dependencies on c and $\log 1/\delta$.

Technically the methods of all five bounds are similar, but some details and much of the analysis differ. Our results as well as the results of [4] use recursive constructions to avoid exponential dependency on c, however the shape of the recursion differs. We show all of our results using the p-moment method, while [4] Theorem 1 and [25, 5] are shown using 2nd-moment analysis. This explains much of why their dependency on δ is worse.

Approximate Kernel Expansions

A classic result by Rahimi and Rect [26] shows how to compute an embedding for any shift-invariant kernel function $k(\|x-y\|_2)$ in time $O(dm)$. In [18] this is improved to any kernel on the form $k(\langle x,y\rangle)$ and time $O((m+d)\log d)$. This is basically optimal in terms of time and space, however the method does not handle kernel functions that can't be specified as a function of the inner product, and it doesn't provide subspace embeddings. See also [22] for more approaches along the same line.

Tensor Sparsification

There is also a literature of tensor sparsification based on sampling [23], however unless the vectors tensored are already very smooth (such as ± 1 vectors), the sampling has to be weighted by the data. This means that these methods in aren't applicable in general to the types of problems we consider, where the tensor usually isn't known when the sketching function is sampled.

Hyper-plane rounding

An alternative approach is to use hyper-plane rounding to get vectors on the form ± 1. Let $\rho = \frac{\langle x,y\rangle}{\|x\|\|y\|}$, then we have $\langle \operatorname{sign}(Mx), \operatorname{sign}(My)\rangle = \sum_i \operatorname{sign}(M_i x)\operatorname{sign}(M_i y) = \sum_i X_i$, where X_i are independent Rademachers with $\mu/m = E[X_i] = 1 - \frac{2}{\pi}\arccos\rho = \frac{2}{\pi}\rho + O(\rho^3)$. By tail bounds then $\Pr[|\langle \operatorname{sign}(Mx), \operatorname{sign}(My)\rangle - \mu| > \epsilon\mu] \le 2\exp(-\min(\frac{\epsilon^2\mu^2}{2\sigma^2}, \frac{3\epsilon\mu}{2}))$. Taking $m = O(\rho^{-2}\epsilon^{-2}\log 1/\delta)$ then suffices with high probability. After this we can simply sample from the tensor product using simple sample bounds.

The sign-sketch was first brought into the field of data-analysis by [6] and [27] was the first, in our knowledge, to use it with tensoring. The main issue with this approach is that it isn't a linear sketch, which hinders some applications, like subspace embeddings. It also takes dm time to calculate Mx and My. In general we would like fast-matrix-multiplication type results.

2 Preliminaries

We will use the following notation

k, i, j	Indicies
c	Tensor order
d	Original dimension (Assumed to be a power of 2.)
d^c	Implicit dimension
m	Sketch dimension
λ	Subspace dimension
Λ	Subspace of R^{d^c}
M	Sketching matrix
$\mathcal{M}$	Distribution of sketching matrices

We say $f(x) \lesssim g(x)$ if $f(x) = O(g(x))$. For $p \ge 1$ and random variables $X \in R$, we write $\|X\|_p = (E|X|^p)^{1/p}$. Note that $\|X+Y\|_p \le \|X\|_p + \|Y\|_p$ by the Minkowski Inequality.

▶ **Definition 4** (Direct sum). *We define the direct sum of two vectors as*

$$x \oplus y = \begin{bmatrix} x \\ y \end{bmatrix},$$

and the direct sum between two matrices as

$$A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}.$$

▶ **Definition 5** (Kronecker (tensor) product). *We define the tensor-product (or Kronecker) of two matrices as:*

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,n}B \\ \vdots & \ddots & \vdots \\ a_{m,1}B & \cdots & a_{m,n}B \end{bmatrix},$$

and in particular of two vectors: $x \otimes y = [x_1y_1, x_1y_2, \dots, x_ny_n]^T$. *Taking the tensor-product of a vector with itself, we get the tensor-powers:*

$$x^{\otimes k} = \underbrace{x \otimes \cdots \otimes x}_{k \text{ times}}.$$

The Kronecker product has the useful mixed-pruduct property when the sizes match up:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

We note in particular the vector variants $(I \otimes B)(x \otimes y) = x \otimes By$ and $\langle x \otimes y, z \otimes t\rangle = \langle x, y\rangle\langle z, t\rangle$.

▶ **Definition 6** (Hadamard product). *Also sometimes known as the 'element-wise product':*

$$x \circ y = [x_1y_1, x_2y_2, \dots, x_ny_n]^T.$$

Taking the Hadamard product with itself gives the Hadamard-power:

$$x^{\circ k} = \underbrace{x \circ \cdots \circ x}_{k \text{ times}} = [x_1^k, x_2^k, \dots, x_n^k]^T.$$

Definitions

▶ **Definition 7** (JL-moment property). *We say a distribution over random matrices* $M \in \mathbb{R}^{m\times d}$ *has the* (ϵ, δ)*-JL-moment property, when*

$$\| \|Mx\|_2^2 - 1\|_p \le \epsilon\delta^{1/p}$$

for all $p > 1$ *and* $x \in \mathbb{R}^d$, $\|x\| = 1$.

Note that by Markov's inequality, the JL-moment-property implies $E\|Mx\|_2 = \|x\|_2$ and that taking $m = O(\epsilon^{-2}\log 1/\delta)$ suffices to have $\Pr[|\|Mx\|_2 - \|x\|_2| > \epsilon] < \delta$ for any $x \in \mathbb{R}^d$. (This is sometimes known as the Distributional-JL property.)

▶ **Definition 8** ((ϵ, δ)-Approximate Matrix Multiplication). *We say a distribution over random matrices* $M \in \mathbb{R}^{k\times d}$ *has the* (ϵ, δ)*-Approximate Matrix Multiplication property if for any matrices* A, B *with proper dimensions,*

$$\begin{aligned} &\| \|A^TM^TMB - A^TB\|_F\|_p \\ &\qquad \le \epsilon\delta^{1/p}\|A\|_F\|B\|_F. \end{aligned}$$

▶ **Lemma 9** (Shown in [28]). *Any distribution that has the* (ϵ, δ)*-JL-moment-property has the* $(3\epsilon, \delta)$*-Approximate Matrix Multiplication property.*

We note that the factor of 3 on ϵ can be removed by combining the analysis in [28] with Appendix Lemma 30.

▶ **Definition 10** (ϵ-Subspace embedding). $M \in \mathbb{R}^{k\times D}$ *is a subspace embedding for* $\Lambda \subseteq \mathbb{R}^D$ *if for any* $x \in \Lambda$,

$$|\|Mx\|_2 - \|x\|_2| \le \epsilon.$$

▶ **Definition 11** ((λ, ϵ)-Oblivious Subspace Embedding). *A distribution,* $\mathcal{M}$, *over* $R^{m\times D}$ *matrices is a* (D, λ)*-Oblivious Subspace Embedding if for any linear subspace,* $\Lambda \subseteq \mathbb{R}^D$, *of dimension* λ, $M \sim \mathcal{M}$ *is an* ϵ*-subspace embedding for* Λ *with probability at least* $1 - \delta$.

▶ **Lemma 12.** *Any distribution that has the* $(\epsilon/(3\lambda), \delta)$*-JL-moment-property is a* (λ, ϵ)*-oblivious subspace embedding.*

Proof. Let $U \in \mathbb{R}^{\lambda\times m}$ be orthonormal such that $U^TU = I$, it then suffices (by [28]) to show $\|U^TM^TMU - I\| \le \epsilon$.

From lemma 9 we have that $\|U^TM^TMU - I\| \le 3\epsilon\delta^{1/p}\|U\|_F^2 = 3\epsilon\delta^{1/p}\lambda$. ◀

▶ **Lemma 13.** *There is a* $C > 0$, *such that any distribution that has the* $(\epsilon, \delta e^{C\lambda})$*-JL-moment-property is a* (λ, ϵ)*-oblivious subspace embedding.*

Proof. For any λ-dimensional subspace, Λ, there exists an ϵ-net $T \subseteq \Lambda \cap S^{d-1}$ of size C^d such that if M preserves the norm of every $x \in T$ then M preserves all of Λ up to $1+\epsilon$. See [28] for details. ◀

▶ **Lemma 14** (Khintchine's inequality [10]). *Let $p \geq 1$, $x \in R^d$, and $\sigma \mathbb{R}^d$ be independent Rademacher ± 1 random variables. Then*

$$\Big\| \sum_{i=1}^{d} \sigma_i x_i \Big\|_p \lesssim \sqrt{p} \|x\|_2.$$

3 Technical Overview

The main component of any tensor sketch is a matrix $M : \mathbb{R}^{m \times d_1 d_2}$ such that $\|Mx\|_2 \approx \|x\|_2$ and which an be applied efficiently (faster than md_1d_2) to simple tensors $x = x^{(1)} \otimes x^{(2)}$, where $x^{(1)} \in \mathbb{R}^{d_1}, x^{(2)} \in \mathbb{R}^{d_2}$.

If $x^{(1)}$ and $x^{(2)}$ are ± 1 vectors, sampling from x works well and can be done without actually constructing x. For this reason a natural general sketch is $S(M^{(1)}x^{(1)} \otimes M^{(2)}x^{(2)})$, where $M^{(1)}$ and $M^{(2)}$ are random rotations.

The original Tensor Sketch did $\mathcal{F}^{-1}(\mathcal{F}C^{(1)}x \circ \mathcal{F}C^{(2)}x)$, where $C^{(1)}$ and $C^{(2)}$ are Count Sketch matrices. At first sight this may look somewhat different, but we can ignore the orthonormal $\mathcal{F}^{-1}$, and then we have $M^{(1)}x^{(1)} \circ M^{(2)}x^{(2)}$ which is just sampling the diagonal of $M^{(1)}x^{(1)} \otimes M^{(2)}x^{(2)}$. Since $M^{(1)}$ and $M^{(2)}$ are independent, sampling the diagonal works as well as any other subset of the same size.

Since Tensor Sketch only used 2nd moment analysis, the natural technical question is "how well does $M^{(1)}x^{(1)} \circ M^{(2)}x^{(2)}$ really work?" when $M^{(1)}$ and $M^{(2)}$ can be anything. In Theorem 27 we show that an embedding dimension of $m = \Theta(\epsilon^{-2} \log 1/\delta + \epsilon^{-1}(\log 1/\delta)^2$ is both sufficient and necessary for (sub)-gaussian matrices, which we conjecture is optimal across all distributions.

Sub-gaussian matrices however still take md time to evaluate, so our tensor sketch would still take $m(d_1 + d_2)$ time in total. We really want $M^{(1)}$ and $M^{(2)}$ to have fast matrix-vector multiplication. It is thus natural to analyze the above scheme where $M^{(1)}$ and $M^{(2)}$ are Fast Johnson Lindenstrauss matrices ala [1, 15]. We do this in Section 5 and show that $m = \epsilon^{-2}(\log 1/\delta)(\log 1/\epsilon\delta)^2$ suffices. For ϵ not too small, this matches our suggested optimum by one $\log 1/\delta$ factor.

The final challenge is to scale up to larger tensors than order 2. Our Lemma 19 shows and exponential dependency: $m = \epsilon^{-2}(\log 1/\delta)(\log 1/\epsilon\delta)^c$, which would be rather unfortunate. Luckily it turns out, that by continuously 'squashing' the dimension back down to $\epsilon^{-2}(\log 1/\delta)(\log 1/\epsilon\delta)^2$, we can avoid this explosion.

While we usually think of applying our sketching matrix to simple tensors, we always analyze everything assuming the input has full rank. This adds some extra difficulty to the analysis, but it is worth it, since by showing that our matrix has the so called JL-moment-property, we get that it is also a subspace embedding for free, by Lemma 12 and Lemma 13.

4 The High Probability Tensor Sketch

In this section we will prove Theorem 3 which is the backbone of our theorems. Theorem 2 will follow as an easy corollary, while Theorem 1 is completed in the next section.

Before we show the full theorem we will consider a slightly easier construction. Given independent random matrices $Q^{(2)}, \ldots, Q^{(c)} \in \mathbb{R}^{m \times dm}$, from a distribution to be discussed later, and $Q^{(1)} \in \mathbb{R}^{m \times d}$, we define $M^{(0)} = 1 \in \mathbb{R}$ and recursively $M^{(k)} = Q^{(k)}(M^{(k-1)} \otimes I_d)$

for $k \in [c]$. The goal of this section is to show that M^k has JL- and related properties when the $Q^{(i)}$s have, and that $M^{(k)}$ can be evaluated efficiently on simple tensors, $x^{(1)} \otimes \ldots \otimes x^{(k)} \in \mathbb{R}^{d^k}$, for $k \in [c]$.

First we show a rather simple fact which will prove to be quite powerful.

▶ **Lemma 15.** *Let $\varepsilon \in (0,1)$ and $\delta > 0$. If $P \in \mathbb{R}^{m_1 \times d_1}$ and $Q \in \mathbb{R}^{m_2 \times d_2}$ are two matrices with (ϵ, δ)-JL moment property, then $P \oplus Q \in \mathbb{R}^{(m_1+m_2)\times(d_1+d_2)}$ has (ϵ,δ)-JL moment property.*

Proof. Let $x \in \mathbb{R}^{d_1+d_2}$ and choose $y \in \mathbb{R}^{d_1}$ and $z \in \mathbb{R}^{d_2}$ such that $x = y \oplus z$. Now using the triangle inequality and JL moment property, we get that

$$\begin{aligned}
\left\|\left\|(P \oplus Q)x\right\|_2^2 - \left\|x\right\|_2^2\right\|_p &\le \left\|\left\|Py\right\|_2^2 - \left\|y\right\|_2^2\right\|_p + \left\|\left\|Qz\right\|_2^2 - \left\|z\right\|_2^2\right\|_p \\
&\le \varepsilon\delta^{1/p}\left\|y\right\|_2^2 + \varepsilon\delta^{1/p}\left\|z\right\|_2^2 \\
&= \varepsilon\delta^{1/p}\left\|x\right\|_2^2,
\end{aligned}$$

since $\|y\|^2 + \|z\|_2 = \|y \oplus z\|_2$ by disjointness. ◀

An easy consequence of this lemma is that for any matrix T, $I_\ell \otimes T$ has (ε, δ)-JL moment property when T has (ε,δ)-JL moment property, since $I_\ell \otimes Q = \underbrace{Q \oplus Q \oplus \ldots \oplus Q}_{\ell \text{ times}}$.

Similarly, $Q \otimes I_\ell$ has (ε,δ)-JL moment property, since you can obtain $Q \otimes I_\ell$ by reordering the rows of $I_\ell \otimes Q$, which trivially does not change the JL moment property.

It is now easy to show that $M^{(k)}$ has JL-property when $Q^{(1)}, \ldots, Q^{(k)}$ has JL-property for $k \in [c]$.

▶ **Lemma 16.** *Let $\varepsilon \in (0,1)$ and $\delta > 0$. If $Q^{(1)}, \ldots, Q^{(c)}$ has the $(\varepsilon/2c, \delta/c)$-JL property, then $M^{(k)}$ has $(k/c\varepsilon, k/c\delta)$-jl property for every $k \in [c]$.*

Proof. Let $k \in [c]$ be fixed. We note that an alternative way of expressing $M^{(k)}$ is as follows:

$$M^{(k)} = Q^{(k)}(Q^{(k-1)} \otimes I_d)(Q^{(k-2)} \otimes I_{d^2}) \ldots (Q^{(1)} \otimes I_{d^{k-1}})$$

Let $x \in \mathbb{R}^{d^k}$ be any vector. Define $x^{(i)} = (Q^{(i)} \otimes I_{d^{k-i}})x^{(i-1)}$ for $i \in [k]$ and $x^{(0)} = x$. Since $Q^{(i)}$ has $(\varepsilon/2k, \delta/k)$-JL property then $Q^{(i)} \otimes I_{d^i}$ has $(\varepsilon/2c,\delta/c)$-JL property by the previous discussion, hence $\Pr\left[\left|\left\|x^{(i)}\right\|_2^2 - \left\|x^{(i)}\right\|_2^2\right| \ge \varepsilon/2c\left\|x^{(i)}\right\|_2^2\right] \le \delta/c$. Now a simple union bound give us that

$$1 - k/c\varepsilon \le (1-\varepsilon/2c)^k \le \left|\left\|Mx\right\|_2^2 - \left\|x\right\|_2^2\right| \le (1+\varepsilon/2c)^k \le 1 + k/c\varepsilon$$

with probability at least $1 - k/c\delta$, which finishes the proof. ◀

▶ **Corollary 17.** *Let $\varepsilon \in (0,1)$. If $Q^{(1)}, \ldots, Q^{(c)}$ has the $(\epsilon/(2\lambda c), \delta/c)$-JL property, then $M^{(k)}$ is a λ-subspace embedding.*

Proof. This follows from lemma 12. ◀

Note that if $Q^{(i)}x$, where $x \in \mathbb{R}^{dk}$, can be evaluated in time T for every $i \in [c] \setminus \{1\}$, and $Q^{(1)}y$, where $y \in \mathbb{R}^d$, also can be evaluated in time T', then $M^{(k)}z$, where $z \in \mathbb{R}^{d^k}$, can be evaluated in time $T'd^{k-1} + T\sum_{i=0}^{k-2} d^i = T'd^{k-1} + T(d^{k-1}-1)/(d-1) = \Theta(T'd^{k-1} + Td^{k-2})$. Meanwhile, if $x \in \mathbb{R}^{d^k}$ is on the form $x^{(1)} \otimes \ldots \otimes x^{(k)}$, we have $M^{(k)}x = Q^{(k)}(M^{(k-1)}(x^{(1)} \otimes$

$\dots x^{(k-1)}) \otimes x^{(k)})$. Now an easy induction argument shows that this allows evaluation in time $O(T' + Tk)$, which is exponentially faster.

Using this construction it now becomes easy to sketch polynomials. More precisely, let $t \in \mathbb{Z}_{>0}$, $k_i \in [c]$ for every $i \in [t]$, then the matrix $M = \bigoplus_{i\in[t]} M^{(k_i)}$ has (ε, δ)-JL property and can be evaluated at the vector $x = \bigoplus_{i\in[t]} \bigotimes_{j\in[k_i]} x^{(i,j)}$ in time $O(T't + T\sum_{i\in[t]} k_i)$, where $x^{(i,j)} \in \mathbb{R}^d$ for every $i \in [t], j \in [k_i]$.

This discussion proves Theorem 3. Note that if we apply a Fast Johnson Lindenstrauss Transform between every direct sum we can obtain an output dimension of $O(m)$.

▶ **Example 18.** Often it is possible to get an even faster evaluation time if the input has even more structure. For example consider the matrix $M = \bigoplus_{i\in[c]} M^{(i)}$ and the vector $z = \bigoplus_{i\in[c]} \bigotimes_{j\in[i]} x^{(j)}$, where $x^{(j)} \in \mathbb{R}^d$ for every $j \in [c]$. Then Mz can be evaluated in time $O(T' + cT)$ by exploiting the fact that

$$M^{(k)}\Big(\bigotimes_{j\in[k]} x^{(j)}\Big) = Q^{(k)}\Big(M^{(k-1)}\Big(\bigotimes_{j\in[k-1]} x^{(j)}\Big) \otimes x^{(k)}\Big) ,$$

so we can use the previous calculations.

As promised we now get the proof of Theorem 2 by choosing $Q^{(1)}, \dots, Q^{(c)}$ to be Fast Johnson Lindenstrauss Matrices. Using the analysis from Krahmer et al. [15] they can be evaluated in time $O(md \log md)$ and if we set $m = \tilde{O}(c^2 \log(1/\delta)/\varepsilon^2)$ then $Q^{(1)}, \dots, Q^{(c)}$ has $(\varepsilon/2c, \delta/c)$-JL property. Now Theorem 3 give us the result.

5 Fast Constructions

The purpose of this section is to show the following lemma:

▶ **Lemma 19.** *Let $c \in \mathbb{Z}_{>0}$, and $(D^{(i)})_{i\in[c]} \in \prod_{i\in[c]} \mathbb{R}^{d_i \times d_i}$ be independent diagonal matrices with independent Rademacher variables. Define $d = \prod_{i\in[c]} d_i$ and $D = \bigotimes_{i\in[c]} D_i \in \mathbb{R}^{d\times d}$. Let $S \in \mathbb{R}^{m\times d}$ be an independent sampling matrix which samples exactly one coordinate per row. Let $x \in \mathbb{R}^d$ be any vector and $p \geq 1$, then*

$$\Big\| \tfrac{1}{m} \|SHDx\|_2^2 - \|x\|_2^2 \Big\|_p \lesssim \sqrt{p}\,(p + \log m)^{c/2} \|x\|_2^2 / \sqrt{m} + p\,(p + \log m)^c \|x\|_2^2 / m.$$

Setting $m = O(\epsilon^{-2} \log 1/\delta (\log 1/\epsilon\delta)^c)$ thus suffices for SHD to have the (ϵ, δ)-JL-moment-property. This then gives (by lemma 9 and 12) that SHD is a subspace embedding.

We note that SHD can be applied efficiently to simple tensors by the relation:

$$\begin{aligned} SH_{d_1 d_2}(D^{(1)} \otimes D^{(2)})(x \otimes y) &= (S^{(1)} \otimes S^{(2)})(H_{d_1} \otimes H_{d_2})(D^{(1)} \otimes D^{(2)})(x \otimes y) \\ &= S^{(1)} H_{d_1} D^{(1)} x \circ S^{(2)} H_{d_2} D^{(2)} y, \end{aligned}$$

where H_n is the size n Hadamard matrix and $S^{(1)}$ and $S^{(2)}$ are independent sampling matrices. Combining this fact with the construction in the previous section gives Theorem 1.

The rest of this section is devoted to proving Lemma 19. We first show two technical lemmas, which seem like they could be useful for many other things.

▶ **Lemma 20.** *Let $p \geq 1$, $c \in Z_{>0}$, and $(\sigma^{(i)})_{i\in[c]} \in \prod_{i\in[c]} \mathbb{R}^{d_i}$ be independent Rademacher vectors. Let $a_{i_0,\dots,i_{c-1}} \in \mathbb{R}$ for every $i_j \in [d_j]$ and every $j \in [c]$, then*

$$\Big\| \sum_{i_1\in[d_1],\dots,i_c\in[d_c]} \prod_{j\in[c]} \sigma^{(j)}_{i_j} a_{i_0,\dots,i_{c-1}} \Big\|_p \lesssim p^{c/2} \Big(\sum_{i_1\in[d_1],\dots,i_c\in[d_c]} a^2_{i_0,\dots,i_{c-1}} \Big)^{1/2} = p^{c/2} \|a\|_{HS}.$$

Proof. The proof will be by induction on c. For $c = 1$ then the result is just Khintchine's inequality (Lemma 14). So assume that the result is true for every value up to c. Using the induction hypothesis we get that

$$\begin{aligned}
\Big\| \sum_{\substack{i_1\in[d_1],\\ \dots,i_c\in[d_c]}} \prod_{j\in[c]} \sigma^{(j)}_{i_j} a_{i_1,\dots,i_c} \Big\|_p &= \Big\| \sum_{\substack{i_1\in[d_1],\\ \dots,i_{c-1}\in[d_{c-1}]}} \prod_{j\in[c-1]} \sigma^{(j)}_{i_j} \Big(\sum_{i_c\in[d_c]} \sigma^{(c)}_{i_c} a_{i_1,\dots,i_c} \Big) \Big\|_p \\
&\lesssim p^{(c-1)/2} \Big\| \Big(\sum_{\substack{i_1\in[d_1],\\ \dots,i_{c-1}\in[d_{c-1}]}} \Big(\sum_{i_c\in[d_c]} \sigma^{(c)}_{i_c} a_{i_1,\dots,i_c} \Big)^2 \Big)^{1/2} \Big\|_p && \text{(I.H.)} \\
&= p^{(c-1)/2} \Big\| \sum_{\substack{i_1\in[d_1],\\ \dots,i_{c-1}\in[d_{c-1}]}} \Big(\sum_{i_c\in[d_c]} \sigma^{(c)}_{i_c} a_{i_1,\dots,i_c} \Big)^2 \Big\|_{p/2}^{1/2} \\
&\le p^{(c-1)/2} \Big(\sum_{\substack{i_1\in[d_1],\\ \dots,i_{c-1}\in[d_{c-1}]}} \Big\| \Big(\sum_{i_c\in[d_c]} \sigma^{(c)}_{i_c} a_{i_1,\dots,i_c} \Big)^2 \Big\|_{p/2} \Big)^{1/2} && \text{(Triangle)} \\
&= p^{(c-1)/2} \Big(\sum_{\substack{i_1\in[d_1],\\ \dots,i_{c-1}\in[d_{c-1}]}} \Big\| \sum_{i_c\in[d_c]} \sigma^{(c)}_{i_c} a_{i_1,\dots,i_c} \Big\|_p^2 \Big)^{1/2} \\
&\lesssim p^{c/2} \Big(\sum_{i_1\in[d_1],\dots,i_c\in[d_c]} a^2_{i_1,\dots,i_c} \Big)^{1/2} && \text{(Khintchine)}
\end{aligned}$$

where the last inequality is by using Khintchine's inequality. Plugging this into the previous inequality finishes the induction step and hence the proof. ◀

The next lemma we nee is a type of Chernoff bound for pth moments.

▶ **Lemma 21.** *Let $p \ge 2$ and $X_0, \dots, X_{k-1}$ be independent non-negative random variables with p-moment, then*

$$\Big\| \sum_{i\in[k]} (X_i - E[X_i]) \Big\|_p \lesssim \sqrt{p} \sqrt{\sum_{i\in[k]} E[X_i]} \Big\| \max_{i\in[k]} X_i \Big\|_p^{1/2} + p \Big\| \max_{i\in[k]} X_i \Big\|_p$$

Proof.

$$\begin{aligned}
\Big\| \sum_{i\in[k]} (X_i - \mathrm{E}[X_i]) \Big\|_p &\lesssim \Big\| \sum_{i\in[k]} \sigma_i X_i \Big\|_p && \text{(Symmetrization)} \\
&\lesssim \sqrt{p} \Big\| \sqrt{\sum_{i\in[k]} X_i^2} \Big\|_p && \text{(Khintchine's inequality)} \\
&= \sqrt{p} \Big\| \sum_{i\in[k]} X_i^2 \Big\|_{p/2}^{1/2} \\
&\le \sqrt{p} \Big\| \max X_i \Big\|_p^{1/2} \Big\| \sum_{i\in[k]} X_i \Big\|_p^{1/2} && \text{(Hölder's inequality)} \\
&\le \sqrt{p} \Big\| \max X_i \Big\|_p^{1/2} \sqrt{\sum_{i\in[k]} \mathrm{E}[X_i]} \\
&\quad + \sqrt{p} \Big\| \max X_i \Big\|_p^{1/2} \Big\| \sum_{i\in[k]} (X_i - \mathrm{E}[X_i]) \Big\|_p^{1/2} && \text{(Triangle inequality)}
\end{aligned}$$

Now let $C = \big\| \sum_{i\in[k]} (X_i - \mathrm{E}[X_i]) \big\|_p^{1/2}$, $B = \sqrt{\sum_{i\in[k]} \mathrm{E}[X_i]}$, and $A = \sqrt{p} \big\| \max X_i \big\|_p^{1/2}$. then we have shown $C^2 \le A(B + C)$. That implies C is smaller than the largest of the roots of the quadratic. Solving this quadratic inequality gives $C^2 \lesssim AB + A^2$ which is the result. ◀

We can finally go ahead and prove Lemma 19.

Proof. For every $i \in [m]$ we let S_i be the random variable that says which coordinate the i'th row of S samples, and we define the random variable $Z_i = M_i x_i = H_{S_i} D x_i$. We note that since the variables $(S_i)_{i\in[m]}$ are independent then the variables $(Z_i)_{i\in[m]}$ are conditionally independent given D, that is, if we fix D then $(Z_i)_{i\in[m]}$ are independent.

Using Lemma 21 we get that

$$\Big\|\frac{1}{m}\sum_{i\in[m]} Z_i^2 - \|x\|_2^2\Big\|_p \lesssim \sqrt{p}\,\Big(\frac{1}{m}\sum_{i\in[m]} \mathrm{E}\big[Z_i^2 \mid D\big]\Big)^{1/2} \Big\|\max_{i\in[m]} \frac{1}{m} Z_i^2\Big\|_p^{1/2} + p\,\Big\|\max_{i\in[m]} \frac{1}{m} Z_i^2\Big\|_p \tag{1}$$

It follows easily that $\mathrm{E}\big[Z_i^2 \mid D\big] = \|x\|_2^2$ from the fact that $\|HDx\|_2^2 = d\|x\|_2^2$, hence $\big(\frac{1}{m}\sum_{i\in[m]} \mathrm{E}\big[Z_i^2 \mid D\big]\big)^{1/2} = \|x\|_2$. Now we just need to bound $\|\max_{i\in[m]} \frac{1}{m} Z_i^2\|_p = \frac{1}{m}\|\max_{i\in[m]} Z_i^2\|_p$. First we note that

$$\|Z_i^2\|_p = \|(H_{S_i} D x_i)^2\|_p \lesssim p^c \|x\|_2^2$$

by Khintchine's inequality. Let $q = \max\{p, \log m\}$, then we get that

$$\Big\|\max_{i\in[m]} Z_i^2\Big\|_p \le \Big\|\max_{i\in[m]} Z_i^2\Big\|_q \le \Big(\sum_{i\in[m]} \|Z_i^2\|_q^q\Big)^{1/q} \le m^{1/q} q^c \|x\|_2^2$$

Now since $q \ge \log m$ then $m^{1/q} \le 2$ so $\|\max_{i\in[m]} Z_i^2\|_p \lesssim q^c\|x\|_2^2 \le (p+\log m)^c \|x\|_2^2$. Plugging this into 1 finishes the proof. ◀

6 Applications

The classic application of Tensor Sketching is compact bilinear pooling (or multilinear). This simply corresponds to expanding x ans $x^{\otimes 2}$ (bilinear pooling) and then hashing back to a smaller size (compact). First discussed in [9] which showed how to do back-propagration through a tensor-sketch layer. Then applied to all many applications such as visual convolutional models [20], question answering [8], visual reasoning [12], video classification [21].

These results are usually given without any particular guarantees, but we can also use polynomial embeddings for concrete algorithms using the following lemma:

6.1 Sketching Polynomials

▶ **Theorem 22.** *Given any degree c polynomial, $P(z) = \sum_{i=0}^{c} a_i z^i$, there are a pair of embeddings $f, g : \mathbb{R}^d \to \mathbb{R}^m$, such that for any $x, y \in \mathbb{R}^d$, the inner product*

$$\langle f(x), g(y)\rangle = (1 \pm \epsilon) P(\langle x, y\rangle)$$

with probability at least $1-\delta$. Using Construction A we set $m = O(c^2\varepsilon^{-2}(\log 1/\delta)(\log 1/\varepsilon\delta)^2)$, and $f(x)$ and $g(y)$ can be computed in $O(c(d\log d + m\log m))$ time. Using Construction B we set $m = \tilde{O}(c^2\varepsilon^{-2}(\log 1/\delta))$, and $f(x)$ and $g(y)$ can be computed in $O(cm\min(d,m))$ time.

Proof. We note that

$$P(\langle x,\, y\rangle) = \sum_{i=0}^{c} a_i \langle x,\, y\rangle^i = \sum_{i=0}^{c} \langle a_i x^{\otimes i},\, y^{\otimes i}\rangle = \Big\langle \bigoplus_{i=0}^{c} a_i x^{\otimes i},\, \bigoplus_{i=0}^{c} y^{\otimes i}\Big\rangle.$$

So using Theorem 3 together with Construction A and Construction B give the result. ◀

We note that the output dimension m in the theorem can be improved by applying a Fast Johnson Lindenstrauss Transform in the end.

▶ **Example 23** (Explicit Gaussian Kernel). Say we want $\langle f(x), g(y)\rangle \approx \exp(-\langle x,y\rangle^2) = \sum_{k=0}^{c}(-1)^k\langle x,y\rangle^{2k}/k! + O(\langle x,y\rangle^{2c+1}/c!)$. Then using Theorem 22 we can obtain $\langle f(x), g(y)\rangle = (1\pm\varepsilon)\sum_{k=0}^{c}(-1)^k\langle x,y\rangle^{2k}/k!$, hence get an approximation of $\varepsilon + O(\langle x,y\rangle^{2c+1}/c!)$ with probability $1-\delta$, using $O(c(d\log d + m\log m))$ time where $m = O(c^3\varepsilon^{-2}(\log 1/\delta)(\log 1/\varepsilon\delta)^2)$.

6.2 Embeddings

▶ **Lemma 24** (Symmetric Polynomials). *Given any degree polynomial, $P(x_1,\ldots,x_d,y_1,\ldots,y_d)$ with k monomials we can make an embedding $f, g : R^d \to R^m$ such that*

$$E\langle f(x), g(y)\rangle = \sum_{\pi} P(x_{\pi(1)},\ldots,x_{\pi(d)},y_{\pi(1)},\ldots,y_{\pi(d)}).$$

$f(x)$ and $g(y)$ can be computed in time $k\,4^{\Delta}$ sketching operations, where Δ is the maximum combined degree of a monomial. (E.g. 4 for $x_1^2 y_1 y_2$.)

For x and y boolean, we get $\|f(x)\|_2^2 \le k(\|x\|_2^2+\kappa-1)!/(\|x\|_2^2-1)! \le k(\|x\|_2^2+\kappa-1)^{\kappa}$. where κ is the numer of different indicies in a monomial. E.g. for $x_1 y_1 x_2$ it is 2.

Proof. See the Appendix Section 7.2. ◀

▶ **Example 25** (Triangle counting). Say you have a database of graphs, $\mathcal{G}$, seen as binary vectors in $\{0,1\}^e$. Given a query graph, G, you want to find $G' \in \mathcal{G}$ such that the number of triangles in $G \cup G'$ is maximized.

We construct the polynomial $P(x,y) = (x_1+y_1-x_1y_1)(x_2+y_2-x_2y_2)(x_3+y_3-x_3y_3)$ which is 1 exactly when $x \cup u$ has a triangle on edges $1,2,3$. The maximum number of different indicies is $\kappa = 3$. The maximum number of triangles (with ordering) is $6\binom{d,3}{\approx}d^3$. We have $\|f(x)\|_2\|g(x)\| \approx (d+3)^3$, so to get precision within 1% of the maximum number, we need to set $\epsilon < 0.01d^3/(d+3)^3$.

Since our approximation works with high probability, we can take a union bound and plug the embedded vectors into the standard data structure of [3] or others.

6.3 Oblivious Subspace Embeddings

In [28] the authors show a number of applications of polynomial kernels in oblivious subspace embeddings. They also show that the original tensor sketch [25] is an oblivious subspace embedding when the sketch matches the size described in the introduction. It is shown how each of:

1. Approximate Kernel PCA and Low Rank Approximation,
2. Regularizing Learning With the Polynomial Kernel,
3. Approximate Kernel Principal Component Regression,
4. Approximate Kernel Canonical Correlation Analysis,

can be computed with state of the art performance.

However, each of the applications encounter an exponential dependency on c. They also inherit the tensor-sketch linear dependency on the inverse error probability. Our sketch improves each of these aspects directly black box.

References

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

2 Josh Alman, Timothy M Chan, and Ryan Williams. Polynomial representations of threshold functions and algorithmic applications. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 467–476. IEEE, 2016.

3 Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 47–66. Society for Industrial and Applied Mathematics, 2017.

4 Anonymous. Sketching high-degree polynomial kernels. In double blind review, 2019.

5 Haim Avron, Huy L. Nguyen, and David P. Woodruff. Subspace embeddings for the polynomial kernel. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2258–2266, 2014. URL: `http://papers.nips.cc/paper/5240-subspace-embeddings-for-the-polynomial-kernel`.

6 Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

7 Michael B Cohen, TS Jayram, and Jelani Nelson. Simple analyses of the sparse johnson-lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

8 Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

9 Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

10 Uffe Haagerup and Magdalena Musat. On the best constants in noncommutative khintchine-type inequalities. *Journal of Functional Analysis*, 250(2):588–624, 2007.

11 Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms (TALG)*, 3(3):31, 2007.

12 Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

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

14 William B Johnson, Gideon Schechtman, and Joel Zinn. Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *The Annals of Probability*, pages 234–253, 1985.

15 Felix Krahmer and Rachel Ward. New and improved johnson–lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

16 Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

17 Rafał Latała et al. Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006.

18 Quoc Viet Le, Tamás Sarlós, and Alexander Johannes Smola. Fastfood: Approximate kernel expansions in loglinear time. *CoRR*, abs/1408.3060, 2014. URL: `http://arxiv.org/abs/1408.3060`, `arXiv:1408.3060`.

19 Joseph Lehec. Moments of the gaussian chaos. In *Séminaire de Probabilités XLIII*, pages 327–340. Springer, 2011.

20 Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

21 Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

22 Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017.

23 Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229, 2015.

24 Rasmus Pagh. Compressed matrix multiplication. *ACM Transactions on Computation Theory (TOCT)*, 5(3):9, 2013.

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

26 Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

27 Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM (JACM)*, 62(2):13, 2015.

28 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014. URL: `https://doi.org/10.1561/0400000060`, `doi:10.1561/0400000060`.

7 Appendix

7.1 Subgaussian construction

Before stating the theorem, we not the following matrix product:

▶ **Definition 26** (Face-splitting product). *Defined between to matrices as the Kronecker-product between pairs of rows:*

$$C \bullet D = \begin{bmatrix} C_1 \otimes D_1 \\ C_2 \otimes D_2 \\ \dots \\ C_n \otimes D_n \end{bmatrix}.$$

Face-splitting product has the relation $(A \bullet B)(x \otimes y) = Ax \circ By$.

▶ **Theorem 27** (Subgaussian). *Let $T, M \in \mathbb{R}^{m \times d}$ have iid. sub-gaussian entries, then $\| \| \frac{1}{\sqrt{m}}(T \bullet M)x\|_2^2 - \|x\|_2^2 \|_p \le \sqrt{p/m} + pq/m$, where $q = \max(p, m)$.*

This immediately implies that for $m = \Omega(\epsilon^{-2} \log 1/\delta + \epsilon^{-1}(\log 1/\delta)(\log 1/\epsilon\delta))$, $T \bullet M$ has the JL-moment property.

We note that the analysis is basically optimal. Assume M and T were iid. Gaussian matrices and $x = e_1'^{\otimes 2}$ were a simple tensor with a single 1 entry. Then $|\|(M \bullet T)x\|_2^2 - \|x\|_2^2| = |\|Mx' \circ Tx'\|_2^2 - 1| \sim |(gg')^2 - 1|$ for $g, g' \in R$ iid. Gaussians. Now $\Pr[(gg')^2 > (1+\epsilon)] \approx \exp(-\min(\epsilon, \sqrt{\epsilon}))$, thus requiring $m = \Omega(\epsilon^{-2} \log 1/\delta + \epsilon^{-1}(\log 1/\delta)^2)$ matching our bound up to a $\log 1/\epsilon$.

Proof. Let $Q = T \bullet M \in \mathbb{R}^{n\times ab}$.

$$\|(T\bullet M)x\|_2^2 = \sum_k ((T\bullet M)_k x)^2 = \sum_k (TU_{(i)})^2_{k,k}$$

where $U_{(i)} = (XM^T)_{(i)} = (M_i X)^T$, when x is seen as a $d\times d$ matrix.

We then have sub-gaussians:

$$E\|U_{(i)}\|_2^2 = E\|M_i X\|_2^2 = \sum_k E(M_i X_{(k)})^2 = \sum_k \|X_{(k)}\|^2 = \|X\|_F^2 = 1$$

$$\|\|U_{(i)}\|_2^2\|_p \le \|\|M_i X^T\|_2^2 - 1\| + 1 \le \sqrt{p}\|X^TX\|_F + p\|X^TX\| + 1 \le p$$

$$\|\sum_k \|U_{(k)}\|_2^4\|_p \le \sqrt{pk} + k + pq^2,$$

The last bound followed from independence and bounded variance of the $\|U_{(k)}\|$s. It is possible to go without this assumption though, suffering a small loss in the final dimension.

We bound

$$\|(T\bullet M)x\|_2^2/k - \|x\|_2^2\|_p \le \|\sum_k (T_k U_{(k)})^2 - \|U\|_F^2\|_p/k \tag{2}$$

$$+ \|\|U\|_F^2/k - \|x\|_2^2\|_p. \tag{3}$$

Bounding (3) follows simply from the JL property of M. Bounding (2) is a bit trickier, and we use the Hanson-Wright inequality:

$$\begin{aligned}\|\sum_k (T_k U_{(k)})^2 - \|U\|_F^2\|_p &\le \Big\|\sqrt{p}\Big(\sum_k \|U_{(k)}U_{(k)}^T\|_F^2\Big)^{1/2} + p\max_k \|U_{(k)}U_{(k)}^T\|\Big\|_p \\ &\le \Big\|\sqrt{p}\Big(\sum_k \|U_{(k)}\|_2^4\Big)^{1/2} + p\max_k \|U_{(k)}\|_2^2\Big\|_p \\ &\le \sqrt{p}\Big\|\sum_k \|U_{(k)}\|_2^4\Big\|_{p/2}^{1/2} + pq. \end{aligned}\tag{4}$$

Here we used the maximum trick from the next section to bound the max term.

A short aside: It would be sweet to split

$$\begin{aligned}\|\sum_k \|U_{(k)}\|_2^4\|_{p/2} &\le \|\max\|U_{(k)}\|_2^2 \sum \|U_{(k)}\|_2^2\|_{p/2} \\ &\le \|\max\|U_{(k)}\|_2^2\|_p \,\|\sum \|U_{(k)}\|_2^2\|_p \\ &\le p\,\|\|U\|_F^2\|_p,\end{aligned}$$

but unfortunately the second factor is[1] $\sqrt{pk}+p$, which means we end up with $p(pk)^{1/4}$ term in (4), which is too much. ($p^{3/4}k^{1/4}$ would have been tolerable.) We'll show how to shave this factor p on the $\sqrt{pk}$ term.

[1] $\|\|U\|_F^2\|_p = \|\sum_j M_j X^T X M_j^T\|_p \le \sqrt{pk}\|X\|_F^2 + p\|X\|^2$.

We instead have to work directly on the fourth powers. We use triangle and Bernstein

$$\begin{aligned}
\|\sum_k \|U_{(k)}\|_2^4\|_{p/2} &\le \Big\|\sum_k \|U_{(k)}\|_2^4 - E\|U_{(k)}\|_2^4\Big\|_{p/2} + \sum_k E\|U_{(k)}\|_2^4 \\
&\le \sqrt{p}\Big(\sum_k (E\|U_{(k)}\|_2^4)^2\Big)^{1/2} + p\|\max\|U_{(k)}\|_2^4\|_{p/2} + \sum_k \|\|U_{(k)}\|_2^2\|_2^2 \\
&\le \sqrt{p}\Big(\sum_k \|\|U_{(k)}\|_2^2\|_4^4\Big)^{1/2} + p\|\max\|U_{(k)}\|_2^2\|_p^2 + 4k \\
&\le \sqrt{p}\Big(\sum_k 4^4\Big)^{1/2} + p\max\|\|U_{(k)}\|_2^2\|_q^2 + k \\
&\le \sqrt{pk} + pq^2 + k.
\end{aligned}$$

Now plugging into (2) and (4) we get

$$\begin{aligned}
\|(T \bullet M)x\|_2^2/k - \|x\|_2^2\|_p &\le (\sqrt{p}\sqrt{\sqrt{pk} + pq^2 + k} + pq)/k + \sqrt{p/k} + p/k \\
&\le \sqrt{p/k} + pq/k.
\end{aligned}$$

Finally we can normalize and plug into the power-Markov inequality:

$$\Pr[\|(T \bullet M)x\|_2^2/k - \|x\|_2^2\|_p \ge \epsilon] \le \max(\sqrt{p/k}, pq/k)^p \epsilon^{-p},$$

which gives that we must take

$$k = \epsilon^{-2}\log 1/\delta + \epsilon^{-1}(\log 1/\delta)^2.$$

◀

Proofs of the lemmas:

▶ **Lemma 28** (Max trick). *Let* $q = \max(p, \log k)$, *then*

$$\|\max X_i\|_p \le e\max\|EX_i\|_q.$$

Proof.

$$\begin{aligned}
\|\max X_i\|_p &\le \|\max X_i\|_q \\
&= (E\max X_i^q)^{1/q} \\
&\le (\sum EX_i^q)^{1/q} \\
&\le (k\max EX_i^q)^{1/q} \\
&\le e(\max EX_i^q)^{1/q} \\
&= e\max\|EX_i\|_q.
\end{aligned}$$

◀

▶ **Lemma 29** (p-norm Bernstein). *For independent variables* X_i,

$$\|\sum_i X_i - E\sum_i X_i\|_p \le \sqrt{p}\,\Big(\sum_i EX_i^2\Big)^{1/2} + p\,\|\max_i X_i\|_p.$$

Proof.

$$
\begin{aligned}
\|\sum_i X_i - E\sum_i X_i\|_p &\le \|\sum_i g_i X_i\|_p && \text{(See notes)}\\
&\le \sqrt{p}\,\|\sum_i X_i^2\|_{p/2}^{1/2}\\
&\le \sqrt{p}\,\big(E\sum_i X_i^2\big)^{1/2} + \sqrt{p}\,\|\sum_i X_i^2 - E\sum_i X_i^2\|_{p/2}^{1/2} && \text{(Triangle)}\\
&\le \sqrt{p}\sigma + \sqrt{p}\,\|\sum_i X_i^2 - E\sum_i X_i^2\|_{p/2}^{1/2}\\
&\le \sqrt{p}\sigma + \sqrt{p}\,\|\sum_i g_i X_i^2\|_{p/2}^{1/2}\\
&\le \sqrt{p}\sigma + \sqrt{p}\,\|\max_i X_i \sum_i g_i X_i\|_{p/2}^{1/2}\\
&\le \sqrt{p}\sigma + \sqrt{p}\,\|\max_i X_i\|_p^{1/2}\|\sum_i g_i X_i\|_p^{1/2}. && \text{(Cauchy)}
\end{aligned}
$$

Now let $Q = \|\sum_i g_i X_i\|_p^{1/2}$ and $K = \|\max_i X_i\|_p$, and we have $Q^2 \le \sqrt{p}\sigma + \sqrt{pK}Q$. Because it's a quadratic form, Q is upper bounded by the larger root of $Q^2 - \sqrt{pK}Q - \sqrt{p}\sigma$. By calculation, $Q^2 \le \sqrt{p}\sigma + pK$, which is the theorem. ◀

7.2 Proof of polynomial lemma

Proof of Lemma 24

Proof. For each monomial $\alpha x^S y^T$ of P, where S and T are multisets : $[d] \to \mathbb{N}$, define $[\varkappa^S y^T]_x$ and $[\varkappa^S y^T]_y$ be two vectors in R^ℓ for some ℓ such that

$$
\langle [\varkappa^S y^T]_x, [\varkappa^S y^T]_y \rangle = \sum_\pi x^{\pi S} y^{\pi T} = \sum_\pi \prod_{i=1}^d x_i^{S_{\pi i}} y_i^{T_{\pi i}}. \tag{5}
$$

Then $f(x)$ and $g(y)$ are simple the sketched concatenation of $\alpha[\varkappa^S y^T]_x$ and $[\varkappa^S y^T]_y$ vectors. (Note that since we get $\epsilon\|f(x)\|_2\|g(y)\|_2$ error, it doesn't matter where we put alpha, or if we split it between f and g.)

We can let $[\varkappa^\emptyset y^\emptyset]_x = [\varkappa^\emptyset y^\emptyset]_y = [1] \in \mathbb{R}^1$ (the single 1 vector) and then define recursively:

$$
\begin{aligned}
[\varkappa^S y^T]_x &= x^{\circ S_i} \otimes [\varkappa^{S\setminus i} y^{T\setminus i}]_x \oplus - \bigoplus_{j\in(S\cup T)\setminus i} [\varkappa^{S\setminus i+\{j:S_i\}} y^{T\setminus i+\{j:T_i\}}]_x\\
[\varkappa^S y^T]_y &= y^{\circ T_i} \otimes [\varkappa^{S\setminus i} y^{T\setminus i}]_y \oplus \bigoplus_{j\in(S\cup T)\setminus i} [\varkappa^{S\setminus i+\{j:S_i\}} y^{T\setminus i+\{j:T_i\}}]_y,
\end{aligned}
$$

where i is any index in $S \cup T$. Here we let $S \setminus i$ be S with i removed, and $S + \{j : S_i\}$ be S with S_i added to S_j. It is clear from Theorem 3 that this construction gives (5).

We note that we can compute the norms by $\|[\varkappa^\emptyset y^\emptyset]\|_2^2 = 1$ and

$$
\|[\varkappa^S y^T]_x\|_2^2 = \|x^{\circ S_i}\|_2^2 \cdot \|[\varkappa^{S\setminus i} y^{T\setminus i}]_x\|_2^2 + \sum_{j\in(S\cup T)\setminus i} \|[\varkappa^{S\setminus i+\{j:S_i\}} y^{T\setminus i+\{j:T_i\}}]_x\|_2^2
$$

and equivalently for y. It does however not seem simple to get a closed form in the general case. In the simple case where x and y are $\{0,1\}^d$ vectors we can however show the simple formula:

$$\|[\chi^S y^T]_x\|_2^2 = \frac{(\|x\|_2^2 - 1 + |S \cup T|)!}{(\|x\|_2^2 - 1)!} \leq (\|x\|_2^2 - 1 + |S \cup T|)^{|S \cup T|}$$

and equivalently for y, (Here S and T are normal sets.)

Since there are only $4^{|}S \cup T|$ many states of $[\chi^S y^T]$ the running time is only that many sketching operations.

◀

7.3 Better Approximate Matrix Multiplication

▶ **Lemma 30.** *For any $x, y \in \mathbb{R}^d$, if S has the (ϵ, δ)-JL-moment-property, ($\|\|Sx\|_2 - \|x\|_2\|_p \leq \epsilon\delta^{1/p}\|x\|_2$), then also*

$$\|(Sx)^T(Sy) - x^T y\|_p \leq \epsilon\delta^{1/p}\|x\|_2\|y\|_2$$

Proof. We can assume by linearity of the norms that $\|x\|_2 = \|y\|_2 = 1$. We then use that $\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^T y$ and $\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2x^T y$.

$$\begin{aligned}
\|(Sx)^T(Sy) - x^T y\|_p &= \|\|Sx + Sy\|_2^2 - \|x + y\|_2^2 - \|Sx - Sy\|_2^2 + \|x - y\|_2^2\|_p/4 \\
&\leq (\|\|S(x + y)\|_2^2 - \|x + y\|_2^2\|_p + \|\|S(x - y)\|_2^2 - \|x - y\|_2^2\|_p)/4 \\
&\leq \epsilon\delta^{1/p}(\|x + y\|_2^2 + \|x - y\|_2^2)/4 \quad \text{(JL property)} \\
&= \epsilon\delta^{1/p}(\|x\|_2^2 + \|y\|_2^2)/2 \\
&\leq \epsilon\delta^{1/p}.
\end{aligned}$$

Combined with the argument in [28] this gives that the JL-moment-property implies Approximate Matrix Multiplication without a factor 3 on ϵ. ◀