Almost Optimal Tensor Sketch

Thomas D. Ahle
ITU
thdy@itu.dk

Jakob B. T. Knudsen
University of Copenhagen
jakn@di.ku.dk

3rd September 2019

Abstract

We construct a matrix $M \in \mathbb{R}^{m \otimes d^c}$ with just $m = O(c\,\lambda\,\varepsilon^{-2}\operatorname{poly}\log 1/\varepsilon\delta)$ rows, which preserves the norm $\|Mx\|_2 = (1 \pm \varepsilon)\|x\|_2$ of all x in any given λ dimensional subspace of $\mathbb{R}^d$ with probability at least $1 - \delta$. This matrix can be applied to tensors $x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ in $O(c\,m\min\{d, m\})$ time – hence the name "Tensor Sketch". (Here $x \otimes y = \operatorname{vec}(xy^T) = [x_1 y_1, x_1 y_2, \ldots, x_1 y_m, x_2 y_1, \ldots, x_n y_m] \in \mathbb{R}^{nm}$.)

This improves upon earlier Tensor Sketch constructions by Pagh and Pham [TOCT 2013, SIGKDD 2013] and Avron et al. [NIPS 2014] which require $m = \Omega(3^c \lambda^2 \delta^{-1})$ rows for the same guarantees. The factors of λ, ε^{-2} and $\log 1/\delta$ can all be shown to be necessary making our sketch optimal up to log factors.

With another construction we get λ times more rows $m = \tilde{O}(c\,\lambda^2\,\varepsilon^{-2}(\log 1/\delta)^3)$, but the matrix can be applied to any vector $x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ in just $\tilde{O}(c\,(d + m))$ time. This matches the application time of Tensor Sketch while still improving the exponential dependencies in c and $\log 1/\delta$.

By reductions in Avron et al. this gives new state of the art algorithms for kernel methods in algorithms such as linear regression and PCA.

Technically, we show two main lemmas: (1) For many Johnson Lindenstrauss (JL) constructions, if $Q, Q' \in R^{m \times d}$ are independent JL matrices, the element-wise product $Qx \circ Q'y$ equals $M(x \otimes y)$ for some $M \in \mathbb{R}^{m \times d^2}$ which is itself a JL matrix. (2) If $M^{(i)} \in \mathbb{R}^{m \times md}$ are independent JL matrices, then $M^{(1)}(x \otimes (M^{(2)}y \otimes \ldots)) = M(x \otimes y \otimes \ldots)$ for some $M \in \mathbb{R}^{m \times d^c}$ which is itself a JL matrix. Combining these two results give an efficient sketch for tensors of any size.

To analyze the two constructions, we give new concentration for products of many JL matrices, as well as a higher order version of Khintchine's inequality, related to the higher order Gaussian chaos analysis by Latała [Annals of Probability 2006].

Contents

1 Introduction

A classical generalization of linear models in statistics are so called "bilinear models" [23]. Instead of training a model on a vector $x = \langle x_1, x_2, \ldots, x_d \rangle$ we do it on $x \otimes x = \langle x_1 x_1, x_1 x_2, \ldots x_1 x_d, \ldots, x_d x_d \rangle$. This allows learning relationships that depend non-linearly on the input. The method extends to higher orders, e.g. $x \otimes x \otimes x = \langle x_1 x_1 x_1, \ldots, x_d x_d x_d \rangle$, but it is clear that the produced vectors will quickly get too large to handle. A natural question is whether we can compute a vector that behaves like $x \otimes x \otimes x$, but has much smaller dimension?

Feature hashing [32] and the Johnson Lindenstrauss transformation [16] are the traditional solutions to this problem. To get from dimension d^3, say, to some much smaller m, we apply a matrix $M \in \mathbb{R}^{m \times d^3}$ where $\langle M(x \otimes x \otimes x), y \rangle \approx \langle (x \otimes x \otimes x), y \rangle$ for any vector $y \in \mathbb{R}^m$. The classic approaches to dimensionality reduction are however too slow, since they require the tensor product to be computed exact before they are applied. A shortcut called "Tensor Sketch" was introduced in [28] which allows $M(x \otimes x \otimes x)$ to be computed with roughly $O(d + m)$ time and space. Random weights applied to tensor products was tested in [10] in combination with convolutional neural networks for classification. The success follows a recent neural network trend in which a layer with random weights often works just as well as one trained for the task at hand [11]. The issue with the approach of [28] and [10] however is that it grows exponentially in the number of tensorings, thus $x \otimes \cdots \otimes x$ (c times) takes time $\Omega(3^c)$ to compute with their approach.

A more rigorous approach was taken by Avron et al. [3] and Woodruff [33]. They proved that a Tensor Sketch with sufficiently many rows is an oblivious "subspace embedding" and used this to give state of the art algorithms for problems such as kernel linear regression and kernel PCA. For example, if we want to take a simple regression model like $\arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2$, where $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^d$, and turn it into a bilinear model, we make a new matrix $X' \in \mathbb{R}^{n \times d^2}$ such that $X'_i = X_i \otimes X_i = \langle X_{i,1} X_{i,1}, X_{i,1} X_{i,2}, \ldots, X_{i,n} X_{i,n} \rangle$, and solve $\arg\min_{w \in \mathbb{R}^{d^2}} \|X'w - y\|_2$. To speed this up, we instead solve $\arg\min_{w \in \mathbb{R}^m} \|(M^T X')w - y\|_2$ for a sketch matrix M. It can be shown that the optimal w will always be in the subspace spanned by the rows of X', so taking M with sufficiently many rows that it preserves all vectors within any subspace of dimension n, the algorithm succeeds with error $\pm\varepsilon$. Taking advantage of the rule $\langle x \otimes x, y \otimes y \rangle = \langle x, y \rangle^2$ we can extend this method to any polynomial kernel function. For any polynomial $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^k$ there are functions $f, g \in \mathbb{R}^d \to \mathbb{R}^m$ such that $\langle f(x), g(y) \rangle = \phi(x, y)$ and using the above method k can be taken to be $O(3^c \varepsilon^{-2} n^2)$ where c is the degree of ϕ. The problem with this method is that some classical kernel functions, such as the Gaussian radial bias function requires c super constant to observe a polynomial approximation. The quadratic dependency on the subspace dimension is also not ideal. In this work we show how to get rid of both of these problems.

In some applications of kernel functions, the algorithms at hand only depend on the inner product between vectors and the "kernel trick" saves having to produce explicit representations $f(x)$ and $g(y)$. However calculating the kernel function ϕ quickly becomes a bottleneck, in particular since the oracle-only nature of the kernel trick tends to require that it is computed for all pairs of data points. There are other methods than tensor

sketch for explicitly sketching vectors for kernels, as we review in our Related Work section, however since they are not linear transformations, they are not known to have the subspace embedding mentioned above, which makes it very hard to get guarantees of correctness. Also note that while the original Tensor Sketch algorithm requires $\Omega(\delta^{-1})$ rows to get success probability $1 - \delta$, one can instead take the median of $\log 1/\delta$ sketches that each succeed with constant probability. However taking the median of multiple sketches is also not a linear operation, and thus for the purpose of subspace embeddings all Tensor Sketch constructions previous to this paper have a huge blow up in the number of rows required to get just $1 - 1/n$ success probability.

1.1 Technical Overview

Say we want a matrix M such that $\|M(x \otimes y)\|_2 = (1 \pm \varepsilon) \|x \otimes y\|_2$ for any $x, y \in \mathbb{R}^d$, and we're willing to spend time most $\tilde{O}(d)$. If x, y where binary $\in \{-1, 1\}^d$ then sampling elements from $x \otimes y$ would have good concentration and so M could simply be a sampling matrix. Generalizing this for $x, y \in \mathbb{R}^d$ we may limit the maximum value by first applying fast random random rotations $R, R' \in \mathbb{R}^{d \times d}$ and then sampling from $Rx \otimes R'y$. By symmetry we might as well sample the diagonal $= Rx \circ R'y = [(Rx)_1(R'y)_1, (Rx)_2(R'y)_2, \ldots, (Rx)_m(R'y)_m]^T$. This has reasonably good concentration by e.g. Bernstein bounds.

What else can be done? The Tensor Sketch of Pham and Pagh [28] computes $C^{(1)}x * C^{(2)}y$, where $C^{(1)}$ and $C^{(2)}$ are independent Count Sketch matrices and $*$ is vector convolution. They show that, amazingly, this equals $C(x \otimes y)$ — a count sketch of the tensor product! This method, surprisingly, turns out the be equivalent to our "first approach", as $C^{(1)}x * C^{(2)}y = \mathcal{F}^{-1}(\mathcal{F}C^{(1)}x \circ \mathcal{F}C^{(2)}y)$, where $\mathcal{F}$ is the Fourier transform. Since $\mathcal{F}$ is an orthonormal matrix, $\mathcal{F}^{-1}$ doesn't impact the norm of Cx and may be ignored. What's left is that Tensor Sketch simply rotates each vector with a matrix $\mathcal{F}C^{(i)}$, similar to the Fast Johnson Lindenstrauss Transform by Ailon et al. [1], and takes the element wise product of the resulting vectors.

Given two matrices $M^{(1)}$ and $M^{(2)}$, the matrix M such that $M(x \otimes y) = M^{(1)}x \circ M^{(2)}y$ is the row-wise tensor product where $M_i = M_i^{(1)} \otimes M_i^{(2)}$ which we write as $M = M^{(1)} \bullet M^{(2)}$. From the above examples it seems like, given any good random rotations $M^{(1)}, M^{(2)}, \ldots$ we can simply construct $M = M^{(1)} \bullet M^{(2)} \bullet \ldots$ and get a good, fast Tensor Sketch. Up to a few issues about independence of the rows of the $M^{(i)}$s we show that this is indeed the case!

By using known random rotations with strong probabilistic guarantees, we show that it suffices to have roughly $(\log 1/\delta)^c$ rows for a cth-order Tensor Sketch. This compares to the second-moment analysis of the Count Sketch based approach, which required $3^c\delta^{-1}$ rows. For small c this is an exponential improvement in $\log 1/\delta$. We would however also like to get rid of the exponential dependence in c. Unfortunately we show a lower bound implying that this dependence is needed for any construction on the form above.

The second idea of this paper is to reduce this dependency to be linear in c, by a recursive construction as follows: After computing a pair-wise sketch $Mx \circ M'y$ we map it back down to a more manageable size, such that $Mx \circ M'(M''y \circ M'''z)$ becomes the sketch of a third order tensor etc. By union bounding the error over c such steps one easily

gets down to a c^2 dependency in the number of rows.

Getting the c dependence down from quadratic to linear takes a much more detailed analysis, which is the main focus of the second section. A key part is the introduction of the following strengthening of the Johnson Lindenstrauss property for random matrices:

Definition 1 ((ε, δ)-Strong JL Moment Property)**.** *Let $\varepsilon, \delta \in [0, 1]$. We say a distribution over random matrices $M \in \mathbb{R}^{m \times d}$ has the (ε, δ)-Strong JL Moment Property, when*
$$E\left[\|Mx\|_2^2\right] = 1 \ and \ \left(E\left[\left(\|Mx\|_2^2 - 1\right)^p\right]\right)^{1/p} \leq \frac{\varepsilon}{e}\sqrt{\frac{p}{\log 1/\delta}}, \ for \ all \ unit \ vectors \ x \in \mathbb{R}^d \ and$$
all p such that $2 \leq p \leq \log 1/\delta$.

The beginning of Section 4 shows more properties of the JL properties and how they relate.

1.2 Theorems

Our most general theorem concerns the recursive combination of matrices for higher order tensor sketching. In stating it we assume that the output dimension, m is between the vector dimension, d and the tensored dimension, d^c. If d is larger than m, one may always start by reducing down to m w.l.o.g.

Theorem 1 (General Sketch)**.** *Assume a distribution over matrices $M \in \mathbb{R}^{m \times dm}$ with the ($\varepsilon/\sqrt{c}, \delta$)-Strong JL Moment Property, and where for any $x \in \mathbb{R}^m, y \in \mathbb{R}^d$, M can be applied to $x \otimes y \in \mathbb{R}^{dm}$ in time T. Then there is a distribution over matrices $M' \in \mathbb{R}^{m \times d^c}$ such that*

1. M' can be applied to vectors on the form $x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ in time $O(cT)$.

2. M' has the ($\varepsilon/\sqrt{c}, \delta$)-Strong JL Moment Property over $\mathbb{R}^{d^c}$.

The theorem also gives results related to Oblivious Subspace Embeddings using Lemma 10 and Lemma 11. The idea is to analyse a matrix with the property $M(x \otimes y \otimes z \dots) = M^{(1)}(x \otimes M^{(2)}(y \otimes M^{(3)}(z \otimes \dots)))$. It turns out that it suffices to show that the Strong JL Moment property is preserved by matrix direct product and multiplication.

Combined with a theorem of Kraemer et al. [19] one may the "Fast JL" matrix of [16] in Theorem 1 to get the two of the results in Table 1. However the direct applications of the theorem doesn't use that we know how to efficiently sketch order 2 tensors. While the number of rows are near optimal, the application time suffers.

To fix this we give two strong families of sketches for low order tensors. For generality we prove them for general c-order tensor products $x^{(1)} \otimes \cdots \otimes x^{(c)}$, but when used with Theorem 1 we will just take $c = 2$.

Theorem 2. *Let $\varepsilon, \delta \in [0, 1]$ and let $c \geq 1$ be some integer. Let $T \in \mathbb{R}^{m \times d}$ be a matrix with iid. rows $T_1, \dots, T_m \in \mathbb{R}^d$ such that $E\left[(T_1 x)^2\right] = \|x\|_2^2$ and $\|T_1 x\|_p \leq \sqrt{ap}\|x\|_2$ for some $a > 0$ and all $p \geq 2$. Let $M = T^{(1)} \bullet \cdots \bullet T^{(c)}$ where $T^{(1)}, \dots, T^{(c)}$ are independent copies of T. Then there exists a constant $K > 0$ such that M has the (ε, δ)-Strong JL Moment Property given*

$$m \geq K \left[(4a)^{2c}\varepsilon^{-2}\log 1/\delta + (2ae)^c\varepsilon^{-1}(\log 1/\delta)^c\right].$$

5

In the case where the $T^{(i)}$ are random Rademacher matrices and c is constant, the bound simplifies to give that $m = O\left(\varepsilon^{-2}\log 1/\delta + \varepsilon^{-1}(\log 1/\delta)^c\right)$ rows suffices. We will show in Appendix B.1 that this analysis is indeed tight.

While the family of Theorem 2 takes advantage of the tensor structure, it is not fast on individual vectors, meaning we still haven't reached our goal of sketching $x^{\otimes c} \in \mathbb{R}^d$ in time near linear in d. We do this in our final construction, which is a version of Fast JL specifically analysed on tensor products:

Theorem 3 (Fast Construction)**.** *Let $T^{(1)} \in \mathbb{R}^{m \times d}$ be a Fast JL matrix, and let $T^{(2)}, \ldots, T^{(c)}$ be independent copies, and define M by $M_i = T_i^{(1)} \otimes \cdots \otimes T_i^{(c)}$. Then taking one can take $m = O(K^c \varepsilon^{-2}(\log 1/\delta)(\log 1/(\varepsilon\delta))^c)$ for some constant K to get*

1. *M has the (ε, δ)-Strong JL Moment Property.*

2. *M can be applied to tensors $x^{(1)} \otimes \cdots \otimes x^{(c)} \in \mathbb{R}^{d^c}$ in time $O(c\,(d\log d + m))$.*

Combined with Theorem 1 these two constructions give the remaining results in Table 1.

We mention a classical application of the above theorems:

Corollary 1 (Polynomial Kernels)**.** *Let $\mathcal{M}$ be a distribution of matrices as in Theorem 1. Let $P : \mathbb{R} \to \mathbb{R}$ be a degree c polynomial, then there is a linear map $M : \mathbb{R}^d \to \mathbb{R}^m$ such that $\langle Mx, My \rangle = P(\langle x, y \rangle)$. M can be computed in time cT.*

This follows simply from the rule $\langle x^{\otimes k}, y^{\otimes k} \rangle = \langle x, y \rangle^k$ combined with Horner's rule $a_0 \oplus a_1 x \oplus a_2 x^{\otimes 2} \oplus \cdots = a_0 \oplus x \otimes (a_1 \oplus x \otimes (a_2 \ldots$. One can extend this technique for general symmetric polynomials $P : \mathbb{R}^{2d} \to \mathbb{R}$ such that $\langle \phi(x), \psi(y) \rangle = P(x, y)$.

1.3 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 [28, 3], since it is most directly comparable to ours. An extra entry in this category is [18], 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 [3], 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 $d \ll m \ll d^c$ 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 an independent JL transformation after embedding. This works because the product of

two subspace embeddings is also a subspace embedding[1], and because standard JL is a subspace embedding by the net-argument (see lemma 11). The embedding dimensions in the table should thus mainly be seen as time and space dependencies, rather than the actual embedding dimension for applications.

Reference	Embedding dimension, m	Embedding time
[28, 3]	$\tilde{O}(3^c \lambda^2 \delta^{-1} \epsilon^{-2})$	$\tilde{O}(c(d+m))$
Theorem 1	$\tilde{O}(c \lambda^2 (\log 1/\delta)^3 \epsilon^{-2})$	$\tilde{O}(c(d+m))$
Theorem 1	$\tilde{O}(c(\lambda + \log 1/\delta)\varepsilon^{-2} + c(\lambda + \log 1/\delta)^2\varepsilon^{-1})$	$\tilde{O}(c\,d\,m)$
Theorem 1	$\tilde{O}(c\lambda^2(\log 1/\delta)\varepsilon^{-2} + c\lambda(\log 1/\delta)^2\varepsilon^{-1})$	$\tilde{O}(c\,d\,m)$
Theorem 1 + [16]	$\tilde{O}(c(\lambda + \log 1/\delta)\varepsilon^{-2})$	$\tilde{O}(c\,d\,m^2)$
Theorem 1 + [19]	$\tilde{O}(c\lambda(\log 1/\delta)^{O(1)}\epsilon^{-2})$	$\tilde{O}(c\,d\,m)$
[18] $(*)$ Theorem 1	$\tilde{O}(c\lambda^2 \delta^{-1} \epsilon^{-2})$	$\tilde{O}(c(d+m))$
[18] $(*)$ Theorem 2	$\tilde{O}(c^6 \lambda (\log 1/\delta)^5 \epsilon^{-2})$	$\tilde{O}(c(d+m))$
[2] $(*), (**)$ Theorem 2.1	$\tilde{O}(\varepsilon^{-2}\log 1/\varepsilon\delta + \varepsilon^{-1}(\log 1/\varepsilon\delta)^c)$	$\tilde{O}(c\,d\,m)$

Table 1: Comparison of embedding dimension and time for some of the results in the article and previous/concurrent work.

A few notes about the contents of Table 1:

- The results with $(*)$ are from unpublished manuscript communicated to us by the authors.

- Some of the results, in particular [28, 3], [18] Theorem 1 and [2] Theorem 2.1 can be applied faster when the input is sparse. Our results, as well as [18], Theorem 2 can similarly be optimized for sparse inputs, by preprocessing vectors with an implementation of Sparse JL [7].

- The result $(**)$ assumes $c = O(1)$ which is why it appears to have fewer dependencies on c.

In comparison to the previous result [28, 3] we are clearly better with an exponential improvement in c as well as δ. Compared to the new work of [18], 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 [18] 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 [18] Theorem 1 and [28, 3] are shown using 2nd-moment analysis. This explains much of why their dependency on δ is worse.

---

[1]Let S and T both preserve subspaces of dimension λ, then given some such subspace, T maps it to some other subspace of dimension at most λ which is then preserved by S.

Approximate Kernel Expansions   A classic result by Rahimi and Rect [29] shows how to compute an embedding for any shift-invariant kernel function $k(\|x-y\|_2)$ in time $O(dm)$. In [22] 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 [24] for more approaches along the same line.

Tensor Sparsification   There is also a literature of tensor sparsification based on sampling [26], 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 $\Pr[|\langle \operatorname{sign}(Mx), \operatorname{sign}(My) \rangle - \mu| > \epsilon\mu] \leq 2\exp(-\min(\frac{\epsilon^2 \mu^2}{2\sigma^2}, \frac{3\epsilon\mu}{2}))$. 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 [31] 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.

1.4   Notation and Preliminaries

Asymptotic notation   We say $f(x) \lesssim g(x)$ if $f(x) \leq Cg(x)$ for all $x \in \mathbb{R}$ and some universal constant C. Note this is slightly different from the usual $f(x) = O(g(x))$ in that it is uniform rather than asymptotic.

For $p \geq 1$ and random variables $X \in R$, we write $\|X\|_p = (E|X|^p)^{1/p}$. Note that $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ by the Minkowski Inequality.

JL Properties   There are a number of different ways to classify JL matrices. The ones we will use are based on the above mentioned moment norm:

Definition 2 (JL-moment property)**.** *We say a distribution over random matrices $M \in \mathbb{R}^{m \times d}$ has the (ε, δ, p)-JL-moment property, when*

$$\big\| \|Mx\|_2^2 - 1 \big\|_p \leq \varepsilon \delta^{1/p}$$

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

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

Note that the Strong JL Moment Property implies the $(\varepsilon, \delta, \log 1/\delta)$-JL Moment Property, since then $\varepsilon\delta^{1/p} = \varepsilon/e$.

Notation for various matrix products   In this article we will be combining matrices in a number of different ways. Here we introduce the definitions and properties thereof which we will use in the later sections.

Tensor product   Given two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{k \times \ell}$ we define the "tensor product" (or Kronecker product), $A \otimes B \in \mathbb{R}^{mk \times n\ell}$ as

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

In particular of two vectors: $x \otimes y = [x_1y_1, x_1y_2, \ldots, x_ny_n]^T$. Notice this equals the flattened outer product xy^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 product property", when the sizes match up: $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. We note in particular the vector variants $(I \otimes B)(x \otimes y) = x \otimes By$, $\langle x \otimes y, z \otimes t \rangle = \langle x, z \rangle \langle y, t \rangle$ and $\langle x^{\otimes k}, y^{\otimes k} \rangle = \langle x, y \rangle^k$.

A related operation is the "direct sum" for vectors: $x \oplus y = \begin{bmatrix} x \\ y \end{bmatrix}$ and for matrices: $A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$. When the sizes match up, we have $(A \oplus B)(x \oplus y) = Ax + By$. Also notice that if I_k is the $k \times k$ identity matrix, then $I_k \otimes A = \underbrace{A \oplus \cdots \oplus A}_{k \text{ times}}$.

Hadamard product   Another useful vector product is the "Hadamard product", also sometimes known as the 'element-wise product'. We define $x \circ y = [x_1y_1, x_2y_2, \ldots, x_ny_n]^T$. Taking the Hadamard product with itself gives the Hadamard-power: $x^{\circ k} = \underbrace{x \circ \cdots \circ x}_{k \text{ times}} = [x_1^k, x_2^k, \ldots, x_n^k]^T$.

Face-splitting product   The Face-splitting product [30] (or transposed Khatri-Rao product) is defined as the rows-by-rows tensor product:

$$M = M^{(1)} \bullet \cdots \bullet M^{(c)} = \begin{bmatrix} M_1^{(1)} \otimes M_1^{(2)} \otimes \cdots \otimes M_1^{(c)} \\ \vdots \\ M_m^{(1)} \otimes M_m^{(2)} \otimes \cdots \otimes M_m^{(c)} \end{bmatrix}.$$

The product has the property (which follows directly from the tensor product), that $(M \bullet T)(x \otimes y) = Mx \circ Ty$.

1.5 Auxiliary Lemmas

In this section we state the auxiliary lemmas that we will use throughout the paper. Most of the lemmas are already known and the proofs of the new lemmas are deferred to Appendix A.

Lemma 1 (Khintchine's inequality [12]). *Let $p > 0$, $x \in \mathbb{R}^d$, and $(\sigma_i)_{i \in [d]}$ be independent Rademacher ± 1 random variables. Then*

$$\left\| \sum_{i=1}^d \sigma_i x_i \right\|_p \leq C_p \, \|x\|_2^2 \,,$$

where $C_p \leq \sqrt{2} \left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}} \right)^{1/p} \leq \sqrt{p}$ for $p \geq 1$.

One may replace (σ_i) with an arbitrary independent sequence of random variables (ς_i) with $E[\varsigma_i] = 0$ and $\|\varsigma_i\|_p \lesssim \sqrt{p}$, and the lemma still holds up to a universal constant factor on the rhs.

We also need a version of Khintchine's inequality that work on tensors of Rademacher vectors.

Lemma 2 (Generalized Khintchine's Inequality). *Let $p \geq 1$, $c \in Z_{>0}$, and $(\sigma^{(i)} \in \mathbb{R}^{d_i})_{i \in [c]}$ be independent vectors each satisfying the Khintchine inequality $\left\| \langle \sigma^{(i)}, x \rangle \right\|_p \leq C_p \|x\|_2$ for any vector $x \in \mathbb{R}^{d_i}$. Let $(a_{i_1, \ldots, i_c} \in \mathbb{R})_{i_j \in [d_j], j \in [c]}$ be a tensor in $\mathbb{R}^{d_1 \times \cdots \times d_c}$, then*

$$\left\| \sum_{i_1 \in [d_1], \ldots, i_c \in [d_c]} \left(\prod_{j \in [c]} \sigma_{i_j}^{(j)} \right) a_{i_1, \ldots, i_c} \right\|_p \leq C_p^c \left(\sum_{i_1 \in [d_1], \ldots, i_c \in [d_c]} a_{i_1, \ldots, i_c}^2 \right)^{1/2}.$$

Or, considering $a \in \mathbb{R}^{d_1 \cdots d_c}$ a vector, then simply $\left\| \langle \sigma^{(1)} \otimes \cdots \otimes \sigma^{(c)}, a \rangle \right\|_p \leq C_p^c \|a\|_2$.

This is related to Latała's estimate for Gaussian chaoses [21], but more simple in the case where a is not assumed to have special structure. Note that this implies the classical bound on the fourth moment of products of 4-wise independent hash functions [5, 15, 27], since $C_4 = 3^{1/4}$ for Rademachers we have $\left\| \langle \sigma^{(1)} \otimes \cdots \otimes \sigma^{(c)}, x \rangle \right\|_4^4 \leq 3^c \, \|x\|_2^4$ for four-independent $\sigma^{(i)}$s.

Proof. The proof is deferred to Appendix A $\qquad \square$

A very useful result for computing the p-norm of a sum of random variables is the following:

Lemma 3 (Latała's inequality, [20]). *If $p \geq 2$ and $X, X_1, \ldots, X_n$ are iid. mean 0 random variables, then we have*

$$\left\| \sum_{i=1}^n X_i \right\|_p \sim \sup \left\{ \frac{p}{s} \left(\frac{n}{p} \right)^{1/s} \|X\|_s \, \middle| \, \max \left\{ 2, \frac{p}{n} \right\} \leq s \leq p \right\}.$$

The following lemma first appeared in [14], but the following is roughly taken from [9], which we also recommend for readers interested in more general versions.

Lemma 4 (General decoupling, [9] Theorem 7.3.1, paraphrasing). *Given a sequence $X_1, \ldots, X_n$ of random variables and a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$. Define $Y_1, \ldots, Y_n$ such that*

1. *$E[Y_i \mid \mathcal{F}_{i-1}] = E[X_i \mid \mathcal{F}_{i-1}]$ for all i.*

2. *The sequence $(Y_i)_i$ is conditionally independent given $X_1, \ldots, X_n$.*

3. *$E[Y_i \mid \mathcal{F}_{i-1}] = E[Y_i \mid X_1, \ldots, X_n]$ for all i.*

Then for all $p \geq 1$,

$$\left\| \sum_i X_i \right\|_p \lesssim \left\| \sum_i Y_i \right\|_p$$

The next lemma is a type of Rosenthal inequality, but which mixes large and small moments in a careful way. It bears similarity to the one sided bound in [4] (Theorem 15.10) derived from the Efron Stein inequality, and the literature has many similar bounds, but we still include a proof here based on first principles.

Lemma 5 (Rosenthal-type inequality). *Let $p \geq 2$ and $X_0, \ldots, X_{k-1}$ be independent non-negative random variables with p-moment, then*

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

Proof. The proof is deferred to Appendix A. □

2   Constructions

The moral of this section is that, if M is a JL matrix, and M' is an independent copy, then for many classic constructions of such matrices, $M \bullet M'$ is also a JL matrix. This is important for tensor sketching, because we have the identity $(M \bullet M')(x \otimes y) = Mx \circ M'y$. That is, matrices on this form can be efficiently applied to tensors.

Note that it is unfortunately not possible to give a general guarantee from the JL property. To see this, note that it doesn't destroy the JL property of a matrix with m rows to add m more rows with all 0s. If we append m such rows to M and prepend as many rows to M' then $M \bullet M' = 0$.

If the rows of M are independent the guarantee does hold. This follows essentially from our analysis in Section 2.1. Unfortunately many interesting JL distributions don't have independent rows. One particular example is the Fast Johnson Lindenstrauss algorithm by Ailon and Chazelle [1]. They considered a matrix $SHD \in \mathbb{R}^{m \times d}$ where $S \in R^{m \times d}$ is a sampling matrix with one 1 per row; $H \in \mathbb{R}^{d \times d}$ is the Hadamard matrix defined as $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes k}$ when $d = 2^k$; and $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix with ± 1 entries. Because all

three matrices allow fast matrix multiplication, Fast JL allows reducing the dimension from d to m in time $O(m + d \log d)$.

We show that nevertheless, if SHD is a Fast JL matrix and $S'H'D'$ is an independent copy, then $SHD \bullet S'H'D'$ is a JL matrix as well. Our analysis is based on Nelson [25] and loses a factor of $\log 1/\delta$ in m compared to the best known method [19] based on the Restricted Isometry Property, but other than that it gets the precise $(\log 1/\delta)^c$ behavior that we expect from fully independent (and slow) JL matrices.

The constructions in this section all follow the "direct composition" paradigm of the original tensor sketch. For this reason they all incur exponential dependencies in c. In the next section we will reduce this to a linear dependency, but we will do so by combining the constructions of this section parameterised with $c = 2$.

2.1   Matrices with independent and identical rows

The classic JL matrix, $M \in \mathbb{R}^{m \times d}$, is a dense matrix with independent Sub Gaussian entries, such as Gaussians or Rademachers (± 1 with even chance.) The formal definition of such a variable is that $\|X\|_p \lesssim \sqrt{p}$.

In this section we analyse our construction on a slightly more general family of matrices. In particular we consider M constructed as follows:

Theorem 2. *Let $\varepsilon, \delta \in [0, 1]$ and let $c \geq 1$ be some integer. Let $T \in \mathbb{R}^{m \times d}$ be a matrix with iid. rows $T_1, \ldots, T_m \in \mathbb{R}^d$ such that $E\big[(T_1 x)^2\big] = \|x\|_2^2$ and $\|T_1 x\|_p \leq \sqrt{ap}\|x\|_2$ for some $a > 0$ and all $p \geq 2$. Let $M = T^{(1)} \bullet \cdots \bullet T^{(c)}$ where $T^{(1)}, \ldots, T^{(c)}$ are independent copies of T. Then there exists a constant $K > 0$ such that M has the (ε, δ)-Strong JL Moment Property given*

$$m \geq K \left[(4a)^{2c} \varepsilon^{-2} \log 1/\delta + (2ae)^c \varepsilon^{-1} (\log 1/\delta)^c \right] .$$

Remark   In the particular case of rows with iid. Rademachers we get $a = \sqrt{3}/4$ (the Khintchine constant from Lemma 1) and

$$m = O \left(3^c \varepsilon^{-2} \log 1/\delta + \varepsilon^{-1} (2.36 \log 1/\delta)^c \right) .$$

The same constants also holds for standard Gaussian random variables. In the case of constant $c = O(1)$ this is simply $O \left(\varepsilon^{-2} \log 1/\delta + \varepsilon^{-1} (\log 1/\delta)^c \right)$. This improves upon the parallel work in [2] which gets $m = \Omega(\varepsilon^{-2} \log 1/\varepsilon\delta + \varepsilon^{-1} (\log 1/\varepsilon\delta)^c)$ in this range.

If we are only interested in the JL property, and not the strong version, it is possible to get $\left(\frac{\log 1/\delta}{c} \right)^c$ in the right term at the cost of an extra factor e^c on the left term.

The proof is based on a generalized Khintchine inequality for tensor products (Lemma 2), as well as the following consequence of a strong result by Latała (Lemma 3):

Corollary 2. *Let $p \geq 2, C > 0$ and $\alpha \geq 1$. Let $(X_i)_{i \in [n]}$ be iid. mean 0 random variables such that $\|X_i\|_p \sim (Cp)^\alpha$, then $\|\sum_i X_i\|_p \lesssim C^\alpha \max\{2^\alpha \sqrt{pn}, (ep)^\alpha\}$.*

We prove this in Appendix B, but for now we focus on the proof of Theorem 2.

Proof. Without loss of generalization we may assume $\|x\|_2 = 1$. We notice that $\left\|\,\|Mx\|_2^2 - 1\right\|_p \le \left\|\frac{1}{m}\sum_i (M_i x)^2 - 1\right\|_p$ is the mean of iid. random variables. Call these $Z_i = (M_i x)^2 - 1$. Then $EZ_i = 0$ and $\|Z_i\|_p = \left\|(M_i x)^2 - 1\right\|_p \lesssim \left\|(M_i x)^2\right\|_p = \|M_i x\|_{2p}^2$ by symmetrization (see e.g. [25]). Now by the assumption $\|T_1 x\|_p \le \sqrt{ap}\|x\|_2 = \sqrt{ap}$, and by Lemma 2, we get that $\|M_i x\|_p = \left\|(T_i^{(1)} \otimes \cdots \otimes T_i^{(c)})x\right\|_p \le (ap)^{c/2}$, and so $\|Z_i\|_p \le (2ap)^c$ for all $i \in [m]$.

We now use Corollary 2 which implies

$$\left\|\frac{1}{m}\sum_i Z_i\right\|_p \le L \max\left\{(4a)^c \sqrt{p/m},\ (2eap)^c/m\right\},$$

for some constant L. Taking $m = (Le)^2 \max\{(4a)^{2c}\varepsilon^{-2}\log 1/\delta, (2ae)^c \varepsilon^{-1}(\log 1/\delta)^c\}$ we get

$$\left\|\frac{1}{m}\sum_i Z_i\right\|_p \le \max\left\{\frac{\varepsilon}{e}\sqrt{\frac{p}{\log 1/\delta}},\ \frac{\varepsilon}{e^2}\left(\frac{p}{\log 1/\delta}\right)^c\right\} \le \frac{\varepsilon}{e}\sqrt{\frac{p}{\log 1/\delta}},$$

where we used $c \ge 1$ and $p \le \log 1/\delta$. $\qquad\square$

We show in Appendix B.1 that the analysis is optimal up to constant factors. This is harder than showing the upper bound, but for the particular case of $c = 2$ the following simple argument gives the intuition:

Assume M and T are 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 the upper bound.

2.2 Fast Constructions

The construction with independent rows can be applied to order c tensors in the time it takes to do c matrix-vector multiplications. The issue is that those each take time md. For large m and d we would like to get this closer to the size of the input. In this section we analyse an approach that takes just $m + d\log d$ time per matrix multiplication.

As mentioned we will analyse $SHD \bullet S'H'D' \bullet \cdots \in \mathbb{R}^{m \times d^c}$ where $S \in R^{m \times d}$ is a sampling matrix with one 1 per row; $H \in \mathbb{R}^{d \times d}$ is the Hadamard matrix defined as $\left[\begin{smallmatrix}1 & 1 \\ 1 & -1\end{smallmatrix}\right]^{\otimes k}$ when $d = 2^k$; and $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix with ± 1 entries. Each of these allow fast matrix multiplication, which immediately gives that SHD can be applied fast.

What is more surprising is that, given the diagonal of D is a tensor product of shorter Rademacher vectors, the SHD construction is particularly applicable to tensor sketching. For an example of this, see Fig. 1 below.

The properties we will use are the following:

1. If S and S' are iid. sampling matrices with independent rows, then $S \bullet S'$ is also a sampling matrix with independent rows. This follows because sampling a random value in $[d^2]$ can be decomposed as sampling $i_1, i_2 \in [d]$ and then taking $i = i_1 d + i_2$. This is exactly what happens when tensoring a row of S with a row of S'.

$SHD(x \otimes y)$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_1\rho_1 & 0 & 0 & 0 \\ 0 & \sigma_1\rho_2 & 0 & 0 \\ 0 & 0 & \sigma_2\rho_1 & 0 \\ 0 & 0 & 0 & \sigma_2\rho_2 \end{bmatrix} \begin{bmatrix} x_1 y_1 \\ x_2 y_1 \\ x_1 y_2 \\ x_2 y_2 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \bullet \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \left(\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \otimes \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \right) \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \otimes \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right)$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \bullet \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \circ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Figure 1: An example of splitting up Hadamard matrices and tensor-diagonal matrices for the Fast Tensor Johnson Lindenstrauss transformation using the mixed product property of tensor products.

2. Since the Hadamard matrix H of size d equals $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes k}$ for some k (remember we assume d is a power of 2), we naturally have that $H \otimes H$ is the Hadamard matrix of size d^2.

3. If D and D' are diagonal matrices with respectively $\sigma \in \mathbb{R}^d$ and $\rho \in \mathbb{R}^d$ on their diagonals, then it is easy to check that $D \otimes D'$ is a diagonal matrix with $\sigma \otimes \rho$ on its diagonal.

From these facts we have that $SHD \bullet S'H'D' = S''HD''$ which is exactly a Fast JL construction, except for D now having a simple tensor on its diagonal instead of a fully independent Rademacher vector. The theorem below will show that, up to some extra log factors, this is not a problem.

Theorem 3 (Fast Tensor Johnson Lindenstrauss). *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 with $\|x\|_2 = 1$ and $p \geq 1$, then*

$$\left\| \tfrac{1}{m} \|SHDx\|_2^2 - 1 \right\|_p \lesssim \sqrt{\frac{pq^c}{m}} + \frac{pq^c}{m},$$

14

where $q = \max\{p, \log m\}$. Setting $m = \epsilon^{-2} \log 1/\delta (K' \log 1/\epsilon\delta)^c$ for some universal constant K', we get that $\frac{1}{\sqrt{m}} SHD$ satisfies the Strong JL Moment Property (Definition 1).

Note that setting $c = 1$, the analysis is very similar to the Fast Johnson Lindenstrauss analysis in [8, 25].

Proof of Theorem 3. 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 $Z_i = M_i x = H_{S_i} Dx$. 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.

We use Lemma 5, the triangle inequality, and Cauchy-Schwartz to get that

$$\left\| \frac{1}{m} \sum_{i \in [m]} Z_i^2 - 1 \right\|_p$$

$$= \left\| \mathrm{E}\left[\left(\frac{1}{m} \sum_{i \in [m]} Z_i^2 - 1 \right)^p \,\Big|\, D \right]^{1/p} \right\|_p \tag{1}$$

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

$$\leq \frac{\sqrt{p}}{m} \left\| \mathrm{E}\left[\left(\max_{i \in [m]} Z_i^2 \right)^p \,\Big|\, D \right]^{1/(2p)} \sqrt{\sum_{i \in [m]} \mathrm{E}\left[Z_i^2 \mid D \right]} \right\|_p + \frac{p}{m} \left\| \max_{i \in [m]} Z_i^2 \right\|_p \tag{3}$$

$$\leq \frac{\sqrt{p}}{m} \left\| \max_{i \in [m]} Z_i^2 \right\|_p^{1/2} \left\| \sum_{i \in [m]} \mathrm{E}\left[Z_i^2 \mid D \right] \right\|_p^{1/2} + \frac{p}{m} \left\| \max_{i \in [m]} Z_i^2 \right\|_p . \tag{4}$$

Here Eq. (1) follows from the definition of the p-norm as well as the law of total expectation, Eq. (2) follows from Lemma 5, Eq. (3) is the triangle inequality, and Eq. (4) uses Cauchy Schwarz: $\|AB\|_p \leq \|A\|_{2p} \|B\|_{2p}$ as well as a few manipulations of norms and powers.

By orthogonality of H we have $\|HDx\|_2^2 = d \|x\|_2^2$ independent of D. Hence

$$\sum_{i \in [m]} \mathrm{E}\left[Z_i^2 \mid D \right] = \sum_{i \in [m]} \|x\|_2^2 = m .$$

To bound $\left\| \max_{i \in [m]} Z_i^2 \right\|_p$ we first use Lemma 2 to show

$$\left\| Z_i^2 \right\|_p = \|H_{S_i} Dx\|_{2p}^2 = \|Dx\|_{2p}^2 \leq p^k \|x\|_2^2 .$$

We then bound the maximum using a sufficiently high powered sum:

$$\left\| \max_{i \in [m]} Z_i^2 \right\|_p \leq \left\| \max_{i \in [m]} Z_i^2 \right\|_q \leq \left(\sum_{i \in [m]} \left\| Z_i^2 \right\|_q^q \right)^{1/q} \leq m^{1/q} q^c \|x\|_2^2 \leq e q^c ,$$

where the last inequality follows from $q \geq \log m$. This gives us that

$$\left\| \frac{1}{m} \sum_{i \in [m]} Z_i^2 - \|x\|_2^2 \right\|_p \lesssim \sqrt{\frac{pq^c}{m}} + \frac{pq^c}{m} ,$$

which finishes the first part of the proof.

To show the Strong JL Moment Property we choose $q = 2e \log m / \delta$ and m such that $m = K \varepsilon^{-2} (\log 1/\delta) q^c \lesssim \varepsilon^{-2} (\log 1/\delta) (K' \log 1/\varepsilon \delta)^c$ for some universal constants K and K'. Then

$$\sqrt{\frac{pq^c}{m}} + \frac{pq^c}{m} \leq \varepsilon \left(\sqrt{\frac{p}{K \log 1/\delta}} + \frac{\varepsilon p}{K \log 1/\delta} \right) \leq \frac{\varepsilon}{e} \sqrt{\frac{p}{\log 1/\delta}}$$

for $p \leq \log 1/\delta$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

3 The High Probability Tensor Sketch

In the previous section we assumed that a tensor sketch had to look like $M = M^{(1)} \bullet M^{(2)} \bullet \cdots \bullet M^{(c)}$ for some matrices $M^{(i)}$, such that $M(x \otimes y \otimes \dots) = M^{(1)} x \circ M^{(2)} y \circ \dots$. This worked well when c was constant, but as we saw it produced a matrix with a number of rows exponential in c.

In this section we instead consider a "sketch and reduce" approach. We'd like to have

$$M(x \otimes y \otimes z \dots) = M^{(1)}(x \otimes M^{(2)}(y \otimes M^{(3)}(z \otimes \dots)))$$

where we assume the vectors are already sufficiently reduced. If $M^{(i)}$ are good tensor sketches for $c = 2$ this combination should be fast and succinct. The matrix M that expands as above on tensors is

$$M = M^{(c)}(M^{(c-1)} \otimes I_d)(M^{(c-2)} \otimes I_{d^2}) \dots (M^{(1)} \otimes I_{d^{c-1}}),$$

or more formally $M = Q^{(c)}$ where $Q^{(0)} = 1 \in \mathbb{R}$ and recursively $Q^{(i)} = M^{(i)}(Q^{(i-1)} \otimes I_d) \in \mathbb{R}^{m \times d^i}$.

In this section we will prove Theorem 1 using this construction. We do this in two steps: 1) We show that each of $M^{(c-i)} \otimes I_{d^i}$ has the (ε, δ)-Strong JL Moment Propertyand 2) That the product of such matrices have it.

The first step follows rather easily from the following lemma:

Lemma 6. *Let $\varepsilon \in (0,1)$, $\delta > 0$ and $p \geq 0$. If $P \in \mathbb{R}^{m_1 \times d_1}$ and $Q \in \mathbb{R}^{m_2 \times d_2}$ are two random matrices (not necessarily independent) such that $\left\| \|Px\|_2^2 - 1 \right\|_p$ and $\left\| \|Qx\|_2^2 - 1 \right\|_p$ are both upper bounded by κ for all $x \in \mathbb{R}^{d_1}$ (resp. $x \in \mathbb{R}^{d_2}$), $\|x\|_2 = 1$; then $\left\| (P \oplus Q)x \right\|_p \leq \kappa$*

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, we get that

$$\left\| \|(P \oplus Q)x\|_2^2 - \|x\|_2^2 \right\|_p = \left\| \|Py\|_2^2 + \|Pz\|_2^2 - \|y\|_2^2 - \|z\|_2^2 \right\|_p$$

16

$$\leq \left\| \|Py\|_2^2 - \|y\|_2^2 \right\|_p + \left\| \|Qz\|_2^2 - \|z\|_2^2 \right\|_p$$
$$\leq \kappa \|y\|_2^2 + \kappa \|z\|_2^2$$
$$= \kappa \|y \oplus z\|_2^2 \,,$$

which is what we want, since $y \oplus z = x$. $\qquad\square$

An easy consequence of this lemma is that for any matrix Q with the (ε, δ)-Strong JL Moment Property, then $I_\ell \otimes Q$ and $Q \otimes I_\ell$ have the (ε, δ)-Strong JL Moment Property too. This follows simply from $I_\ell \otimes Q = \underbrace{Q \oplus Q \oplus \ldots \oplus Q}_{\ell \text{ times}}$ and the fact that you can obtain $Q \otimes I_\ell$ by reordering the rows and columns of $I_\ell \otimes Q$, which does not change the JL moment property since it corresponds to permutations of the input or output vectors.

We continue to show a "product lemma" for JL matrices, to complement the one we just proved for sums.

Lemma 7. *There exists a universal constant L, such that for any $\varepsilon, \delta \in [0, 1]$, if $M^{(1)} \in \mathbb{R}^{d_2 \times d_1}, \ldots, M^{(c)} \in \mathbb{R}^{d_{c+1} \times d_c}$ are independent random matrices with the Strong $(\varepsilon/(L\sqrt{k}), \delta)$-JL Moment Property, then the matrix $M = M^{(c)} \cdots M^{(1)}$ has the (ε, δ)-Strong JL Moment Property.*

Proof. Let $x \in \mathbb{R}^d$ be an arbitrary, fixed unit vector, and fix $1 < p \leq \log(1/\delta)$. We define $X_i = \left\| M^{(i)} \cdots M^{(1)} x \right\|_2^2$ and $Y_i = X_i - X_{i-1}$ for every $i \in [k]$. By telescoping we then have that $X_k - 1 = \sum_{i \in [k]} Y_i$.

We will prove by induction on $k \in [c]$ that

$$\left\| \sum_{i \in [k]} Y_i \right\|_p \leq \frac{\varepsilon}{e} \sqrt{\frac{t}{\log(1/\delta)}} \leq 1 \quad \text{for all } k \in [c]. \tag{5}$$

Intuitively this should be true, since Y_i's are a sub-Gaussian Martingale difference sequence, however the lack of a moment generating function makes it more tricky to show. Note that since $\sum_{i \in [c]} Y_i = X_c - 1 = \|Mx\|_2^2 - 1$, this is exactly the statement that M has the (ε, δ)-Strong JL Moment Property.

Induction start: For $i = 1$ we have $Y_1 = X_1 - X_0 = \left\| M^{(1)} x \right\|_2^2 - \|x\|_2^2$. Now as $M^{(1)}$ has the $(\varepsilon/(L\sqrt{k}), \delta)$-Strong JL Moment Property we get that

$$\left\| \sum_{i \in [1]} Y_i \right\|_p = \left\| \left\| M^{(1)} x \right\|_2^2 - 1 \right\|_p \leq \frac{\varepsilon}{eL\sqrt{k}} \sqrt{\frac{p}{\log(1/\delta)}} \leq \frac{\varepsilon}{e} \sqrt{\frac{p}{\log(1/\delta)}}.$$

For the induction step we introduce $(T^{(i)})_{i \in [k]}$ as independent copies of $(M^{(i)})_{i \in [k]}$ and define $Z_i = \left\| T^{(i)} M^{(i-1)} \ldots M^{(1)} x \right\|_2^2 - \left\| M^{(i-1)} \ldots M^{(1)} x \right\|_2^2$ for every $i \in [k]$. We can verify the following three properties:

1. $\Pr\!\left[Z_k > t \mid (M^{(j)})_{j \in [k-1]} \right] = \Pr\!\left[Y_k > t \mid (M^{(j)})_{j \in [k-1]} \right]$ for every $t \in \mathbb{R}, k \in [c]$.

17

2. The sequence $(Z_i)_{i\in[c]}$ is conditionally independent given $(M^{(i)})_{i\in[c]}$.

3. $\Pr\left[Z_k > t \mid (M^{(i)})_{i\in[c-1]}\right] = \Pr\left[Z_k > t \mid (M^{(i)})_{i\in[c]}\right]$ for every $t \in \mathbb{R}, k \in [c]$.

This means we can use Lemma 4 to get

$$\left\|\sum_{i\in[k]} Y_i\right\|_p \le C_0 \left\|\sum_{i\in[k]} Z_i\right\|_p \qquad \text{for every } i \in [c] \tag{6}$$

and so it suffices to show Eq. (5) on $\sum_{i\in[k]} Z_i$ which is somewhat more well behaved than $\sum_{i\in[k]} Y_i$.

Now assume that (5) is true for $i-1$. Using (6) we get that $\|X_i - 1\|_p = \left\|\sum_{j\in[i]} Y_j\right\|_p \le C_0 \left\|\sum_{j\in[i]} Z_j\right\|_p$. By using that $(T^{(j)})_{j\in[i]}$ has the Strong $(\varepsilon/(L\sqrt{k}), \delta)$-JL Moment Property together with Khintchine's inequality (Lemma 1), we get that

$$\left\|\sum_{j\in[i]} Z_j\right\|_p = \left\|\mathrm{E}\left[\left(\sum_{j\in[i]} Z_j\right)^p \middle| (M^{(j)})_{j\in[i]}\right]^{1/p}\right\|_p$$

$$\le C_1 \left\|\frac{\varepsilon}{eL\sqrt{k}}\sqrt{\frac{p}{\log(1/\delta)}}\sqrt{\sum_{j\in[i]} X_j^2}\right\|_p$$

$$\le C_1 \frac{\varepsilon}{e}\sqrt{\frac{p}{\log(1/\delta)}}\frac{1}{L\sqrt{k}}\sqrt{\sum_{j\in[i]} \|X_j\|_p^2},$$

where the last inequality follows from the triangle inequality. Using the triangle inequality and (5) we have $\|X_i\|_p \le 1 + \|X_k - 1\|_p = 1 + \left\|\sum_{i\in[k]} Y_i\right\|_p \le 2$ by the induction hypothesis. Setting $L = 2C_0C_1$ we get that

$$\left\|\sum_{i\in[k]} Y_i\right\|_p \le \frac{\varepsilon}{e}\sqrt{\frac{p}{\log(1/\delta)}}\frac{C_0C_1}{L\sqrt{k}}\sqrt{\sum_{i\in[k]} \|X_i\|_p^2}$$

$$\le \frac{\varepsilon}{e}\sqrt{\frac{p}{\log(1/\delta)}}\frac{C_0C_1}{L\sqrt{k}} 2\sqrt{k}$$

$$\le \frac{\varepsilon}{e}\sqrt{\frac{p}{\log(1/\delta)}},$$

which finishes the induction. $\qquad\square$

Combining the two lemmas finally gives Theorem 1.

4   Oblivious Subspace Embedding

In this last section of paper we show how to get an Oblivious Subspace Embedding using our results. First we formally define Oblivious Subspace Embedding.

Definition 3 (ε-Subspace embedding). *$M \in \mathbb{R}^{k \times D}$ is a subspace embedding for $\Lambda \subseteq \mathbb{R}^D$ if for any $x \in \Lambda$,*

$$|\|Mx\|_2 - \|x\|_2| \leq \varepsilon \|x\|_2.$$

Definition 4 (($\lambda, \varepsilon, \delta$)-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 $1 - \delta$.*

We show two different ways to obtain an Oblivious Subspace Embedding. The first approach is using Approximate Matrix Multiplication, which is formally defined below, and the approach is using a standard net-argument.

Definition 5 ((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,*

$$\|\|A^T M^T M B - A^T B\|_F\|_p \leq \varepsilon \delta^{1/p} \|A\|_F \|B\|_F.$$

We will show that the JL-moment property implies Approximate Matrix Multiplication. To do this we first show that the JL-moment property implies concentration on the inner product of two vectors.

Lemma 8 (Two vector JL-moment property). *For any $x, y \in \mathbb{R}^d$, if S has the (ε, δ)-JL-moment-property, then also*

$$\|(Sx)^T(Sy) - x^T y\|_p \leq \varepsilon \delta^{1/p} \|x\|_2 \|y\|_2 \tag{7}$$

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 $\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2x^T y$ such that $x^T y = (\|x + y\|_2^2 - \|x - y\|_2^2)/4$.

Plugging this into the left hand side of Eq. (7) gives

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

$\square$

Now we can prove the following lemma which was first proved in [17] theorem 6.2, see also [33].

Lemma 9 (JL implies Approximate Matrix Multiplication)**.** *Any distribution that has the (ε, δ, p)-JL-moment-property has the (ε, δ, p)-Approximate Matrix Multiplication property.*

We note that [33] has a factor of 3 on ε, but this can be removed with Lemma 8. We give a short proof for completion:

Proof.

$$
\begin{aligned}
\left\| \left\| A^T M^T M B - A^T B \right\|_F \right\|_p &= \left\| \sum_{i,j} (A_i M^T M B_j^T - A_i B_j^T)^2 \right\|_{p/2}^{1/2} \\
&\le \sqrt{\sum_{i,j} \left\| A_i M^T M B_j^T - A_i B_j^T \right\|_p^2} \\
&\le \sqrt{\sum_{i,j} (\varepsilon \delta^{1/p} \left\| A_i \right\|_2 \left\| B_j \right\|_2)^2} \\
&= \varepsilon \delta^{1/p} \left\| A \right\|_F \left\| B \right\|_F .
\end{aligned}
$$

$\square$

Lemma 10. *Any distribution that has the $(\varepsilon/\lambda, \delta)$-JL-moment-property is a (λ, ε)-oblivious subspace embedding.*

Proof. Let $U \in \mathbb{R}^{\lambda \times m}$ be orthonormal such that $U^T U = I$, it then suffices (by [33]) to show $\|U^T M^T M U - I\| \le \varepsilon$. From Lemma 9 we have that $\|U^T M^T M U - I\| \le \varepsilon \delta^{1/p} \|U\|_F^2 = \varepsilon \delta^{1/p} \lambda$. $\square$

The second way to get an Oblivious Subspace Embedding is using a standard net-argument.

Lemma 11. *There is a $C > 0$, such that any distribution that has the $(\varepsilon, \delta e^{-C\lambda})$-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 $1 + \varepsilon$. See [33] for details. $\square$

4.1 Combination lemmas

It is also possible to show lemmas similar to Lemma 6 and Lemma 7 for Oblivious Subspace Embeddings. Combined with an OSE analysis of JL, merging that of [8] and Theorem 3 one gets another way to achieve a fast OSE tensor sketch with linear dependence on λ.

Lemma 12 (A Direct Sum is an OSE)**.** *Let $M \in \mathbb{R}^{m \times d}$ be an $(\varepsilon, \delta/b)$-oblivious subspace embedding, then $M \otimes I_b \in \mathbb{R}^{mb \times db}$ is an (ε, δ)-oblivious subspace embedding.*

Intuitively, let $A_{[i]}$ denote the ith block of $A = [A_{[1]}, \dots, A_{[b]}]^T$. By a union bound M is a subspace embedding for each block. We then intuitively have

$$
\|(M \otimes I_b) A x\|_2^2 = \sum_{i \in [b]} \left\| M A_{[i]} x \right\|_2^2 = (1 \pm \varepsilon) \sum_{i \in [b]} \left\| A_{[i]} x \right\|_2^2 = (1 \pm \varepsilon) \left\| A x \right\|_2^2 .
$$

Proof. We can make this more formal, using that $\|X\| = \sup_x |x^T X x|$ for symmetric X and x a unit vector. This means that the subspace property is equivalent with that $|x^T (A_{[i]}^T M^T M A_{[i]} - A_{[i]}^T A_{[i]}) x| \le \varepsilon \, x^T A_{[i]}^T A_{[i]} x$ for all $i \in [b]$ and vectors x.

Hence

$$\left\| A^T (M \otimes I_b)^T (M \otimes I_b) A - A^T A \right\| = \sup_x \left| x^T (A^T (M \otimes I_b)^T (M \otimes I_b) A - A^T A) x \right|$$

$$= \sup_x \left| \sum_{i \in [b]} x^T (A_{[i]}^T M^T M A_{[i]} - A_{[i]}^T A_{[i]}) x \right|$$

$$\le \sup_x \sum_{i \in [b]} \left| x^T (A_{[i]}^T M^T M A_{[i]} - A_{[i]}^T A_{[i]}) x \right|$$

$$\le \sup_x \sum_{i \in [b]} \varepsilon \, x^T A_{[i]}^T A_{[i]} x$$

$$= \varepsilon \sup_x x^T A^T A x$$

$$= \varepsilon \left\| A^T A \right\|$$

$$= \varepsilon \left\| A \right\|^2 .$$

This implies the non-symmetric case

$$\left\| A^T (M \otimes I_b)^T (M \otimes I_b) B - A^T B \right\| \le 2\varepsilon \left\| A \right\| \left\| B \right\|$$

by an argument in [8]. (Basically taking a single space large enough to contain both A and B.) □

Note that this matches what we have for the JL-moment-property except that it requires a union bound, rather than just individual moment bounds.

Lemma 13 (A Product is an OSE). *Let $M^{(1)}, \dots, M^{(c)}$ have the Moment-OSE property, in other words assume that $\left\| \left| \left\| A^T M^{(i)T} M^{(i)} A - I \right\| \right| \right\|_p \le \varepsilon \delta^{1/p}$ for some p and all i, when $A^T A = I$.*

Then

$$\left\| \left| \left\| A^T M^{(c)T} \cdots M^{(1)T} M^{(1)} \cdots M^{(c)} A - I \right\| \right| \right\|_p \le (1 + \varepsilon \delta^{1/p})^c - 1.$$

Note that for small ε or δ this is $\approx n \varepsilon \delta^{1/p}$.

Proof. We show this for the case $c = 2$.

$$\left\| \left| \left\| A^T (MT)^T (MT) A - I \right\| \right| \right\|_p \le \left\| \left| \left\| A^T T^T M^T M T A - A^T T^T T A \right\| \right| \right\|_p + \left\| \left| \left\| A^T T^T T A - I \right\| \right| \right\|_p$$

$$\le \varepsilon \delta^{1/p} \left\| \left| \left\| A^T T^T T A \right\| \right| \right\|_p + \varepsilon \delta^{1/p}$$

$$\le \varepsilon \delta^{1/p} (\left\| \left| \left\| A^T T^T T A - I \right\| \right| \right\|_p + 1) + \varepsilon \delta^{1/p}$$

$$\le (1 + \varepsilon \delta^{1/p})^2 - 1.$$

By induction these results can be extended to give $(1 + \varepsilon\delta^{1/p})^c - 1 \approx c\varepsilon\delta^{1/p}$ for composing n subspace embeddings.

Alternatively we can show the same thing directly without normalization:

$$\left\|\left\|A^T(MT)^T(MT)B - A^TB\right\|\right\|_p \leq \left\|\left\|A^TT^TM^TMTB - A^TT^TTB\right\|\right\|_p + \left\|\left\|A^TT^TTB - A^TB\right\|\right\|_p$$

$$\leq \varepsilon\delta^{1/p}\left\|\|TA\|\,\|TB\|\right\|_p + \varepsilon\delta^{1/p}\,\|A\|\,\|B\|$$

$$\leq \varepsilon\delta^{1/p}\left\|\|T\|^2\right\|_p\|A\|\,\|B\| + \varepsilon\delta^{1/p}\,\|A\|\,\|B\|$$

$$\leq \varepsilon\delta^{1/p}(\left\|\left\|T^TT - I\right\|\right\|_p + 1)\,\|A\|\,\|B\| + \varepsilon\delta^{1/p}\,\|A\|\,\|B\|$$

$$\leq ((1 + \varepsilon\delta^{1/p})^2 - 1)\,\|A\|\,\|B\|\,.$$

$\square$

It is likely that a "strong OSE moment" property would be able to reduce that to $\approx \sqrt{n}\varepsilon\delta^{1/p}$, similar to the JL case. However this seems to be outside of current random matrix techniques.

5   Acknowledgements

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] Anonymous. High probability bounds for sketching tensor products. In double blind review, 2019.

[3] 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.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[5] Vladimir Braverman, Kai-Min Chung, Zhenming Liu, Michael Mitzenmacher, and Rafail Ostrovsky. AMS without 4-wise independence on product domains. In *27th International Symposium on Theoretical Aspects of Computer Science, STACS 2010, March 4-6, 2010, Nancy, France*, pages 119–130, 2010. URL: https://doi.org/10.4230/LIPIcs.STACS.2010.2449, doi:10.4230/LIPIcs.STACS.2010.2449.

[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] Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[9] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence.* Springer Science & Business Media, 2012.

[10] 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.

[11] Olivia Guest and Bradley C Love. What the success of brain imaging implies about the neural code. *Elife*, 6:e21397, 2017.

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

[13] Paweł Hitczenko. Domination inequality for martingale transforms of a rademacher sequence. *Israel Journal of Mathematics*, 84(1-2):161–178, 1993.

[14] Pawel Hitczenko. On a domination of sums of random variables by sums of conditionally independent ones. *The Annals of Probability*, pages 453–468, 1994.

[15] Piotr Indyk and Andrew McGregor. Declaring independence via the sketching of sketches. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 737–745. Society for Industrial and Applied Mathematics, 2008.

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

[17] Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

[18] Michael Kapralov, Rasmus Pagh, Ameya Velingker, David P. Woodruff, and Amir Zandieh. Sketching high-degree polynomial kernels. In review, 2019.

[19] 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.

[20] Rafał Latała. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997.

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

[22] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, 2013.

[23] 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.

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

[25] Jelani Nelson. Lecture notes on algorithms for big data, October 2015.

[26] 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.

[27] Mihai Patrascu and Mikkel Thorup. The power of simple tabulation hashing. *J. ACM*, 59(3):14:1–14:50, 2012. URL: https://doi.org/10.1145/2220357.2220361, doi:10.1145/2220357.2220361.

[28] 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.

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

[30] VI Slyusar. Generalized face-products of matrices in models of digital antenna arrays with nonidentical channels. *RADIO ELECTRONICS AND COMMUNICATIONS SYSTEMS C/C OF IZVESTIIA-VYSSHIE UCHEBNYE ZAVEDENIIA RADIOELEKTRONIKA*, 46(10):9–17, 2003.

[31] 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.

[32] Kilian Q. Weinberger, Anirban Dasgupta, John Langford, Alexander J. Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1113–1120, 2009. URL: https://doi.org/10.1145/1553374.1553516, doi:10.1145/1553374.1553516.

[33] 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`.

A Proof of Auxiliary Lemmas

Lemma 2 (Generalized Khintchine's Inequality). *Let $p \geq 1$, $c \in Z_{>0}$, and $(\sigma^{(i)} \in \mathbb{R}^{d_i})_{i \in [c]}$ be independent vectors each satisfying the Khintchine inequality $\left\| \langle \sigma^{(i)}, x \rangle \right\|_p \leq C_p \|x\|_2$ for any vector $x \in \mathbb{R}^{d_i}$. Let $(a_{i_1,\ldots,i_c} \in \mathbb{R})_{i_j \in [d_j], j \in [c]}$ be a tensor in $\mathbb{R}^{d_1 \times \cdots \times d_c}$, then*

$$\left\| \sum_{i_1 \in [d_1],\ldots,i_c \in [d_c]} \left(\prod_{j \in [c]} \sigma_{i_j}^{(j)} \right) a_{i_1,\ldots,i_c} \right\|_p \leq C_p^c \left(\sum_{i_1 \in [d_1],\ldots,i_c \in [d_c]} a_{i_1,\ldots,i_c}^2 \right)^{1/2}.$$

Or, considering $a \in \mathbb{R}^{d_1 \cdots d_c}$ a vector, then simply $\left\| \langle \sigma^{(1)} \otimes \cdots \otimes \sigma^{(c)}, a \rangle \right\|_p \leq C_p^c \|a\|_2$.

Proof. The proof will be by induction on c. For $c = 1$ then the result is by assumption. So assume that the result is true for every value up to $c-1$. Let $B_{i_1,\ldots,i_{c-1}} = \sum_{i_c \in [d_c]} \sigma_{i_c}^{(c)} a_{i_1,\ldots,i_c}$. We then pull it out of the left hand term in the theorem:

$$\left\| \sum_{i_1 \in [d_1],\ldots,i_c \in [d_c]} \left(\prod_{j \in [c]} \sigma_{i_j}^{(j)} \right) a_{i_1,\ldots,i_c} \right\|_p = \left\| \sum_{i_1 \in [d_1],\ldots,i_{c-1} \in [d_{c-1}]} \left(\prod_{j \in [c-1]} \sigma_{i_j}^{(j)} \right) B_{i_1,\ldots,i_{c-1}} \right\|_p$$

$$\leq C_p^{c-1} \left\| \left(\sum_{i_1 \in [d_1],\ldots,i_{c-1} \in [d_{c-1}]} B_{i_1,\ldots,i_{c-1}}^2 \right)^{1/2} \right\|_p \quad (8)$$

$$= C_p^{c-1} \left\| \sum_{i_1 \in [d_1],\ldots,i_{c-1} \in [d_{c-1}]} B_{i_1,\ldots,i_{c-1}}^2 \right\|_{p/2}^{1/2}$$

$$\leq C_p^{c-1} \left(\sum_{i_1 \in [d_1],\ldots,i_{c-1} \in [d_{c-1}]} \left\| B_{i_1,\ldots,i_{c-1}}^2 \right\|_{p/2} \right)^{1/2} \quad (9)$$

$$= C_p^{c-1} \left(\sum_{i_1 \in [d_1],\ldots,i_{c-1} \in [d_{c-1}]} \left\| B_{i_1,\ldots,i_{c-1}} \right\|_p^2 \right)^{1/2}.$$

Here Eq. (8) is the inductive hypothesis and Eq. (9) is the triangle inequality. Now $\left\| B_{i_1,\ldots,i_{c-1}} \right\|_p^2 \leq C_p^2 \sum_{i_c \in [d_c]} a_{i_1,\ldots,i_c}^2$ by Khintchine's inequality, which finishes the induction step and hence the proof. □

Lemma 5 (Rosenthal-type inequality). *Let $p \geq 2$ and $X_0, \ldots, X_{k-1}$ be independent*

non-negative random variables with p-moment, then

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

Proof.

$$\left\| \sum_{i \in [k]} (X_i - \mathrm{E}[X_i]) \right\|_p \lesssim \left\| \sum_{i \in [k]} \sigma_i X_i \right\|_p \quad \text{(Symmetrization)}$$

$$\lesssim \sqrt{p} \left\| \sum_{i \in [k]} X_i^2 \right\|_{p/2}^{1/2} \quad \text{(Khintchine's inequality)}$$

$$\leq \sqrt{p} \left\| \left(\max_{i \in [k]} X_i \right) \sum_{i \in [k]} X_i \right\|_{p/2}^{1/2} \quad \text{(Non-negativity)}$$

$$\leq \sqrt{p} \left\| \max X_i \right\|_p^{1/2} \left\| \sum_{i \in [k]} X_i \right\|_p^{1/2} \quad \text{(Cauchy)}$$

$$\leq \sqrt{p} \left\| \max X_i \right\|_p^{1/2} \left(\sqrt{\sum_{i \in [k]} \mathrm{E}[X_i]} + \left\| \sum_{i \in [k]} (X_i - \mathrm{E}[X_i]) \right\|_p^{1/2} \right).$$

Now let $C = \left\| \sum_{i \in [k]} (X_i - \mathrm{E}[X_i]) \right\|_p^{1/2}$, $B = \sqrt{\sum_{i \in [k]} \mathrm{E}[X_i]}$, and $A = \sqrt{p} \left\| \max X_i \right\|_p^{1/2}$. then we have shown $C^2 \leq A(B + C)$. 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. $\qquad \square$

B Upper and Lower Bounds for Sub-Gaussians

The following simple corollaries will be used for both upper and lower bounds:

Corollary 3. *Let* $2n \geq p \geq 2, C > 0$ *and* $\alpha \geq 1$. *Let* $(X_i)_{i \in [n]}$ *be iid. mean 0 random variables such that* $\|X_i\|_p \sim (Cp)^\alpha$, *then*

$$\left\| \sum_i X_i \right\|_p \sim C^\alpha \max\{2^\alpha \sqrt{pn},\, (n/p)^{1/p} p^\alpha\}.$$

Proof. We show that the expression $f(s) = \frac{p}{s} \left(\frac{n}{p} \right)^{1/s} s^\alpha$ in Lemma 3 is maximized either by minimizing or maximizing the parameter s. It suffices to show that the derivative $\frac{df(s)}{ds} \frac{p}{s} =$

$\frac{-p}{s^{3-\alpha}}(\frac{n}{p})^{1/s}\left((1-\alpha)s+\log\frac{n}{p}\right)$ is non-decreasing in s, which follows from $(n/p)^{1/s}s^{\alpha-3}$ being positive. Hence we either take $s=\max\{2,n/p\}$ or $s=p$, but from the assumption $2n\geq p$ that gives the corollary. $\qquad\square$

On the upper bound side we can simplify things a bit.

Corollary 2. *Let $p\geq 2, C>0$ and $\alpha\geq 1$. Let $(X_i)_{i\in[n]}$ be iid. mean 0 random variables such that $\|X_i\|_p\sim(Cp)^\alpha$, then*

$$\left\|\sum_i X_i\right\|_p\lesssim C^\alpha\max\{2^\alpha\sqrt{pn},(ep)^\alpha\}.$$

Proof. The requirement on the upper bound disappears since we can always take the supremum over a larger range without decreasing the result. We get rid of the term $(n/p)^{1/p}\leq n^{1/p}$ with the following argument: Assume $n^{1/p}p^\alpha$ dominates the maximum. Then $2^\alpha\sqrt{pn}\leq n^{1/p}p^\alpha\implies n^{1/p}\leq((p/2)^\alpha p^{-1/2})^{\frac{2}{p-2}}$. One can check that $p^{\frac{-1}{p-2}}\leq 1$ and $(p/2)^{\frac{2}{p-2}}$ is decreasing for $p\geq 2$ and so $n^{1/p}$ is bounded by $\lim_{p\to 2}(p/2)^{\frac{2\alpha}{p-2}}=e^\alpha$. $\qquad\square$

For the lower bound we will also use the following result by Hitczenko, which provides an improvement on Khintchine for Rademacher random variables.

Lemma 14 (Sharp bound on Rademacher sums [13])**.** *Let $\sigma\in\{-1,1\}^n$ be a random Rademacher sequence and let $a\in\mathbb{R}^n$ be an arbitrary real vector, then*

$$\|\langle a,\sigma\rangle\|_p\sim\sum_{i\leq p}a_i+\sqrt{p}\Big(\sum_{i>p}a_i^2\Big)^{1/2}$$

Finally the lower bound will use the Paley-Zygmund inequality (also known as the one-sided Chebyshev inequality):

Lemma 15 (Paley-Zygmund)**.** *Let $X\geq 0$ be a real random variable with finite variance, and let $\theta\in[0,1]$, then*

$$\Pr[X>\theta E[X]]\geq(1-\theta)^2\frac{E[X]^2}{E[X^2]}.$$

A classical strategy when using Paley-Zygmund is to prove $E[X]\geq 2\varepsilon$ and $E[X]^2/E[X^2]>4\delta$ for some $\varepsilon,\delta>0$, and then take $\theta=1/2$ to give $\Pr[X>\varepsilon]>\delta$.

B.1   Lower Bound for Sub-Gaussians

The following lower bound considers the sketching matrix consisting of the direct composition of matrices with Rademacher entries. Note however that the assumptions on Rademachers are only used to show that the p-norm of a single row with a vector is $\sim\sqrt{p}$. For this reason the same lower bound hold if the Rademacher entries are substituted for say Gaussians.

Theorem 4 (Lower bound). *For some constants $C_1, C_2, B > 0$, let $d, m, c \geq 1$ be integers, let $\varepsilon \in [0, 1]$ and $\delta \in [0, 1/16]$. Further assume that $d \geq \log 1/\delta \geq c/B$. Then the following holds.*

Let $M^{(1)}, \ldots, M^{(c)} \in \mathbb{R}^{m \times d}$ be matrices with all independent Rademacher entries and let $M = \frac{1}{\sqrt{m}} M^{(1)} \bullet \cdots \bullet M^{(c)}$. Then there exists some unit vector $y \in \mathbb{R}^{d^c}$ such that if

$$m < C_1 \max \left\{ 3^c \varepsilon^{-2} \frac{\log 1/\delta}{c}, \ \varepsilon^{-1} \left(\frac{C_2 \log 1/\delta}{c} \right)^c \right\} \quad then \quad \Pr \left[\left| \|My\|_2^2 - 1 \right| > \varepsilon \right] > \delta.$$

Proof. Let $y = [1, \ldots, 1]^T / \sqrt{d} \in \mathbb{R}^d$ and let $x = y^{\otimes c}$. We have

$$\|My\|_2^2 - 1 = \frac{1}{m} \left\| M^{(1)} x \circ \cdots \circ M^{(c)} x \right\|_2^2 - 1 = \frac{1}{m} \sum_{j \in [m]} \left(\prod_{i \in [c]} Z_{i,j}^2 \right) - 1$$

where each $Z_{i,j} = \sum_{k \in [d]} M_{j,k}^{(i)} / \sqrt{d}$ are independent averages of d independent Rademacher random variables. By Lemma 14 we have $\|Z_{i,j}\|_p \sim \min\{\sqrt{p}, \sqrt{d}\}$ which is $\sqrt{p}$ by the assumption $d \geq \log 1/\delta$ as long as $p \leq \log 1/\delta$. By the expanding $Z_{i,j}^4$ into monomials and linearity of expectation we get $\|Z_{i,j}\|_4 = \frac{1}{\sqrt{d}} (d + 3d(d-1))^{1/4} = (3 - 2/d)^{1/4}$.

Now define $X_j = \prod_{i \in [c]} Z_{i,j}^2 - 1$, then $E X_j = 0$ and $\|X_j\|_p \geq \left\| \prod_{i \in [c]} Z_{i,j}^2 \right\|_p - 1 = \|Z_{i,j}\|_{2p}^{2c} - 1 \geq K^c p^c$ for some K, assuming $p \geq 2$. In particular, $\|X_j\|_2 \geq \|Z_{i,j}\|_4^{2c} - 1 = (3 - 2/d)^{c/2} - 1 \sim 3^{c/2}$ by the assumption $d \geq c \geq 1$.

We have $\left\| \|My\|_2^2 - 1 \right\|_p = \frac{1}{m} \left\| \sum_{j \in [m]} X_m \right\|_p$ is a sum of iid. random variables, so we can use Corollary 3 to show

$$K_3 \max \left\{ \sqrt{3^c p/m}, (m/p)^{1/p} K_1^c p^c / m \right\} \lesssim \left\| \|My\|_2^2 - 1 \right\|_p$$

$$\lesssim K_4 \max \left\{ \sqrt{3^c p/m}, (m/p)^{1/p} K_2^c p^c / m \right\} \quad (10)$$

for some universal constants $K_1, K_2, K_3, K_4 > 0$.

Assume now that $m < \max \left\{ A K_3^2 3^c \varepsilon^{-2} \frac{\log 1/\delta}{c}, \frac{K_3}{4} \varepsilon^{-1} \left(4 A K_1 \frac{\log 1/\delta}{c} \right)^c \right\}$ as in the theorem. We take $p = 4A \frac{\log 1/\delta}{c}$ for some constant A to be determined. We want to show $\left\| \|My\|_2^2 - 1 \right\|_p \geq 2\varepsilon$. For this we split into two cases depending on which term of $m < \max\{(1), (2)\}$ dominates. If $(1) \geq (2)$ we pick the first lower bound in Eq. (10) and get $\left\| \|My\|_2^2 - 1 \right\|_p \geq K_3 \sqrt{3^c p/m} \geq K_3 \sqrt{\frac{4\varepsilon^2}{K_3^2}} = 2\varepsilon$. Otherwise, if $(2) \geq (1)$, we pick the other lower bound and also get:

$$\left\| \|My\|_2^2 - 1 \right\|_p \geq K_3 (m/p)^{1/p} \frac{K_1^c p^c}{m} \geq \frac{K_3}{2} \frac{K_1^c \left(4A \frac{\log 1/\delta}{c} \right)^c}{\frac{K_3}{4} \varepsilon^{-1} \left(4A K_1 \frac{\log 1/\delta}{c} \right)^c} = 2\varepsilon,$$

where we used $(m/p)^{1/p} \geq e^{-1/(em)} \geq 1/2$ for $m \geq 1$. Plugging into Paley-Zygmund (Lemma 15) we have

$$\Pr \left[\left| \|My\|_2^2 - 1 \right| \geq \varepsilon \right] \geq \Pr \left[\left| \|My\|_2^2 - 1 \right|^p \geq \left\| \|My\|_2^2 - 1 \right\|_p^p 2^{-p} \right]$$

$$\geq \frac{1}{4} \left(\frac{\left\| \|My\|_2^2 - 1 \right\|_p}{\left\| \|My\|_2^2 - 1 \right\|_{2p}} \right)^{2p}, \tag{11}$$

where we used that $p \geq 1$ so $(1 - 2^{-p})^2 \geq 1/4$.

There are again two cases depending on which term of the upper bound in Eq. (10) dominates. If $\sqrt{3^c p/m} \geq (m/p)^{1/p} K_2^c p^c/m$ we have using the first lower bound that $\frac{\left\| \|My\|_2^2 - 1 \right\|_p}{\left\| \|My\|_2^2 - 1 \right\|_{2p}} \geq \frac{K_3}{\sqrt{2}K_4}$. For the alternative case, $(m/p)^{1/p} K_2^c p^c/m \geq \sqrt{3^c p/m}$, we have

$$\frac{\left\| \|My\|_2^2 - 1 \right\|_p}{\left\| \|My\|_2^2 - 1 \right\|_{2p}} \geq \frac{K_3}{\sqrt{2}K_4} \frac{(m/p)^{1/p}}{(m/2p)^{1/2p}} \left(\frac{K_1}{2K_2} \right)^c \geq \frac{K_3}{2K_4} \left(\frac{K_1}{2K_2} \right)^c$$

where $\frac{(m/p)^{1/p}}{(m/2p)^{1/2p}} \geq e^{-1/(4em)} \geq 1/\sqrt{2}$ for $m \geq 1$.

Comparing with (11) we see that it suffices to take $A \leq \min\{\frac{1}{\log 2K_4/K_3}, \frac{1}{\log 2K_2/K_1}\}/32$. This choice also ensures that $1 \leq p \leq \log 1/\delta$ as we promised. Note that we may assume in Eq. (10) that $K_3 \leq K_4$ and $K_1 \leq K_2$. We then finally have

$$\frac{1}{4} \left(\frac{K_3}{\sqrt{2}K_4} \right)^{2p} \geq \frac{1}{4} \delta^{1/(4c)} \quad \text{and} \quad \frac{1}{4} \left(\frac{K_3}{2K_4} \left(\frac{K_1}{2K_2} \right)^c \right)^{2p} \geq \frac{1}{4} \delta^{1/(4c)+1/4},$$

which are both $> \delta$ for $c \geq 1$ and $\delta < 1/16$. $\qquad\square$