Electronic Colloquium on Computational Complexity, Report No. 98 (2010)

A Derandomized Sparse Johnson-Lindenstrauss Transform

Daniel M. Kane† Jelani Nelson‡

Abstract

Recent work of [Dasgupta-Kumar-Sarlós, STOC 2010] gave a sparse Johnson-Lindenstrauss transform and left as a main open question whether their construction could be efficiently derandomized. We answer their question affirmatively by giving an alternative proof of their result requiring only bounded independence hash functions. Furthermore, the sparsity bound obtained in our proof is improved. The main ingredient in our proof is a spectral moment bound for quadratic forms that was recently used in [Diakonikolas-Kane-Nelson, CoRR abs/0911.3389].

1 Introduction

The Johnson-Lindenstrauss lemma states the following.

Lemma 1 (JL Lemma [16]). *For any integers $k, d > 0$, and any $0 < \varepsilon < 1/2$, there exists a probability distribution on $k \times d$ real matrices for $k = \Theta(\varepsilon^{-2} \log(1/\delta))$ such that for any $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$,*

$$\mathbf{Pr}_A[|\|Ax\|_2^2 - 1| > \varepsilon] < \delta.$$

Several proofs of the JL lemma exist in the literature [1, 7, 10, 13, 15, 16], and it is known that the dependence on k is tight up to an $O(\log(1/\varepsilon))$ factor [5]. Though, these proofs of the JL lemma give a distribution over *dense* matrices, where each column has at least a constant fraction of its entries being non-zero, and thus naïvely performing the matrix-vector multiplication is costly. Recently, [9] proved the JL lemma where each matrix in the support of their distribution only has α non-zero entries per column, for $\alpha = \Theta(\varepsilon^{-1} \log(1/\delta) \log^2(k/\delta))$. This reduces the time to perform dimensionality reduction from the naïve $O(dk)$ to being $O(d\alpha)$.

The construction of [9] involved picking two random hash functions $h : [d\alpha] \to [k]$ and $\sigma : [d\alpha] \to \{-1, 1\}$, and thus required $\Omega(d\alpha \cdot \log k)$ bits of seed to represent a random matrix from their JL distribution. They then left two main open questions: (1) derandomize their construction to require fewer random bits to select a random JL matrix, for applications in e.g. streaming settings where storing a long random seed is prohibited, and (2) understand the dependence on δ that is required in α.

We give an alternative proof of the main result of [9] that yields progress for both (1) and (2) above simultaneously. Specifically, our proof yields a value of α that is improved by a $\log(k/\delta)$ factor. Furthermore, our proof only requires that h be r_h-wise independent and σ be r_σ-wise independent for $r_h = O(\log(k/\delta))$ and $r_\sigma = O(\log(1/\delta))$, and thus a random sparse JL matrix can be represented using only $O(\log(k/\delta) \log(d\alpha + k)) = O(\log(k/\delta) \log d)$ bits (note k can be assumed less than d, else the JL lemma is trivial, in which case also $\log(d\alpha) = O(\log d)$). We remark that [9]

[1]Harvard University, Department of Mathematics. dankane@math.harvard.edu.
[2]MIT Computer Science and Artificial Intelligence Laboratory. minilek@mit.edu.

ISSN 1433-8092

asked exactly this question: whether the random hash functions used in their construction could be replaced by functions from bounded independence hash families. The proof in [9] required use of the FKG inequality [6, Theorem 6.2.1], and they suggested that one approach to a proof that bounded independence suffices might be to prove some form of this inequality under bounded independence. Our approach is completely different, and does not use the FKG inequality at all. Rather, the main ingredient in our proof is a spectral moment bound for quadratic forms recently used in [11].

2 Other Related Work

There have been two separate lines of related work: one line of work on constructing JL families[1] such that the dimensionality reduction can be performed quickly, and another line of work on derandomizing the JL lemma so that a random matrix from some JL family can be selected using few random bits. We discuss both here.

2.1 Works on efficient JL embeddings

Here and throughout, for a JL family $\mathcal{A}$ we use the term *embedding time* to refer to the running time required to perform a matrix-vector multiplication for an arbitrary $A \in \mathcal{A}$. The first work to give a JL family with embedding time potentially better than $O(dk)$ was in [2]. There, the authors achieved embedding time $O(d \log d + k \log^2(1/\delta))$. Later, improvements were given by Ailon and Liberty in [3, 4]. The work of [3] achieves embedding time $O(d \log k)$ when $k = O(d^{1/2-\gamma})$ for an arbitrarily small constant $\gamma > 0$, and [4] achieves embedding time $O(d \log d)$ and no restriction on k, though the k in their JL family is $O(\varepsilon^{-4} \log(1/\delta) \log^4 d)$ as opposed to the $O(\varepsilon^{-2} \log(1/\delta))$ bound of the standard JL lemma. Liberty, Ailon, and Singer [18] achieve embedding time $O(d)$ when $k = O(d^{1/2-\gamma})$, but their JL family only applies for x satisfying $\|x\|_\infty \le \|x\|_2 \cdot k^{-1/2} d^{-\gamma}$. None of these works however can take advantage of the situation when x is sparse to achieve faster embedding time. In both [9] and the current work however, if x has support size $\|x\|_0$, we achieve embedding time $O(\|x\|_0 \cdot \alpha)$.

Other related works include [8] and [22]. Implicitly in [8], and more explicitly in [22], it was shown that the JL family of [9] can achieve arbitrary constant error probability $\delta > 0$ as long as h is pairwise independent and σ is 4-wise independent. The claim fails for subconstant δ though, since with such mild independence assumptions on h, σ one needs k to be polynomially large in $1/\delta$.

2.2 Works on derandomizing the JL lemma

Karnin, Rabani, and Shpilka [17] recently gave a JL family where the distortion ε and failure probability δ are $1/k^C$ for some absolute constant $C > 0$ — note that in Lemma 1, the failure probability decays exponentially in $\varepsilon^2 k$. Other works giving derandomized JL lemmas are [11, 19], which give pseudorandom generators (PRGs) against degree-2 polynomial threshold functions (PTFs) over the hypercube. A degree-t PTF is a function $f : \{-1,1\}^d \to \{-1,1\}$ which can be represented as the sign of a degree-t d-variate polynomial. A PRG that δ-fools degree-t PTFs is a

[1] In many known proofs of the JL lemma, the distribution over matrices in Lemma 1 is obtained by picking a matrix uniformly at random from some set $\mathcal{A}$. In such a case, we call $\mathcal{A}$ a *JL family*.

function $F : \{-1,1\}^s \to \{-1,1\}^d$ such that for any degree-t PTF f,

$$|\mathbf{E}_{z\in\mathcal{U}^s}[f(F(z))] - \mathbf{E}_{x\in\mathcal{U}^d}[f(x)]| < \delta,$$

where $\mathcal{U}^m$ is the uniform distribution on $\{-1,1\}^m$.

Note that the conclusion of the JL lemma can be rewritten as

$$\mathbf{E}_A[I_{[1-\varepsilon,1+\varepsilon]}(\|Ax\|_2^2)] \geq 1-\delta,$$

where $I_{[a,b]}$ is the indicator function of the interval $[a,b]$, and furthermore A can be taken to have random ± 1 entries [1]. Noting that $I_{[a,b]}(z) = (\mathrm{sign}(z-a) - \mathrm{sign}(z-b))/2$ and using linearity of expectation, we see that any PRG which δ-fools $\mathrm{sign}(p(x))$ for degree-t polynomials p must also δ-fool $I_{[a,b]}(p(x))$. Now, for fixed x, $\|Ax\|_2^2$ is a degree-2 polynomial over the boolean hypercube in the variables $A_{i,j}$ and thus a PRG which δ-fools degree-2 PTFs also gives a JL family with the same seed length. Each of [11, 19] thus give JL families with seed length $\mathrm{poly}(1/\delta)\cdot\log d$. It can be shown via the probabilistic method that there exist PRGs for degree-2 PTFs with seed length $O(\log(d/\delta))$ (see Section B of the full version of [19] for a proof); an explicit construction of such a PRG would achieve the holy grail for derandomized JL family constructions.

Other derandomizations of the JL lemma include the works [12] and [20]. A common application of the JL lemma is the case where there are n vectors $x_1,\ldots,x_n \in \mathbb{R}^d$ and one wants to find a matrix $A \in \mathbb{R}^{k\times d}$ to preserve $\|x_i - x_j\|_2$ to within relative error ε for all i,j. In this case, one can set $\delta = 1/n^2$ and apply the JL lemma, then perform a union bound over all i,j pairs. The works of [12, 20] do not give JL families, but rather give derandomizations for this application in the case that the vectors $x_1,\ldots,x_n$ are known up front.

3 Conventions and Notation

Definition 2. *For $A \in \mathbb{R}^{n\times n}$, we define the* Frobenius norm *of A as $\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$.*

Definition 3. *For $A \in \mathbb{R}^{n\times n}$, we define the* operator norm *of A as*

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2.$$

In the case A has all real eigenvalues (e.g. it is symmetric), we also have that $\|A\|_2$ is the largest magnitude of an eigenvalue of A.

Throughout this paper, ε is the quantity given in Lemma 1, and is assumed to be smaller than some absolute constant $\varepsilon_0 > 0$. All logarithms are base-2 unless explicitly stated otherwise.

4 Our Main Theorem

We recall the sparse JL transform construction of [9] (though the settings of some of our constants differ). Let $k = 16\cdot 64^2\cdot\varepsilon^{-2}\log(1/\delta)$.[2] Pick random hash functions $h : [d] \to [k]$ and $\sigma : [d] \to \{-1,1\}$. Let $\delta_{i,j}$ be a random variable indicating $h(i) = j$. Define the matrix $A \in \{-1,1\}^{k\times d}$ by

[2]Though our constant factor is quite large, most likely the 64 could be made much smaller by tightening the analysis of constants in [11, Theorem 5.1].

$A_{i,j} = \delta_{i,j} \cdot \sigma(i)$. The work of [9] showed that as long as $x \in \mathbb{R}^d$ satisfies $\|x\|_2 = 1$ and has bounded $\|x\|_\infty$, then $\mathbf{Pr}_{h,\sigma}[|\|Ax\|_2^2 - 1| > \varepsilon] < O(\delta)$. We show the same conclusion without the assumption that h, σ are perfectly random; in particular, we show that h need only be r_h-wise independent and σ need only be r_σ-wise independent for $r_h = O(\log(k/\delta))$ and $r_\sigma = O(\log(1/\delta))$. Furthermore, our assumption on the bound for $\|x\|_\infty$ is $\|x\|_\infty \le c$ for $c = \Theta(\sqrt{\varepsilon/(\log(1/\delta) \cdot \log(k/\delta))})$, whereas [9] required $c = \Theta(\sqrt{\varepsilon/(\log(1/\delta) \cdot \log^2(k/\delta))})$. This is relevant since the column sparsity obtained in the final JL transform construction of [9] is $1/c^2$. This is because, to apply the dimensionality reduction of [9] to an arbitrary x of unit ℓ_2 norm (which might have $\|x\|_\infty \gg c$), one should first map x to a vector $\tilde{x}$ by a $(d/c^2) \times d$ matrix Q with $Q_{i_1 r + i_2, i_1+1} = c$ and other entries 0 for $i_1 \in \{0, \ldots, d-1\}$, $i_2 \in [1/c^2]$. Then $\|\tilde{x}\|_2 = 1$ and $\|\tilde{x}\|_\infty \le c$, and thus the set of products with Q of JL matrices in the distribution of [9] over dimension d/c^2 serves as a JL family for arbitrary unit vectors. Thus, the sparsity obtained by our proof in the final JL construction is improved by a $\Theta(\log(k/\delta))$ factor.

Before proving our main theorem, first we note that

$$\|Ax\|_2^2 = \|x\|_2^2 + 2 \sum_{(s,t) \in \binom{[d]}{2}} \left(\sum_{j=1}^{k} \delta_{s,j} \delta_{t,j} x_s x_t \right) \sigma(s)\sigma(t).$$

We would like that $\|Ax\|_2^2$ is concentrated about 1, or rather, that

$$Z = 2 \sum_{s<t} \left(\sum_{j=1}^{k} \delta_{s,j} \delta_{t,j} x_s x_t \right) \sigma(s)\sigma(t) \tag{1}$$

is concentrated about 0. Let $\eta_{s,t}$ be a random variable indicating that $s \ne t$ and $h(s) = h(t)$. Then note that, for fixed h, Z is a quadratic form in the $\sigma(i)$ which can be written as $\sigma^T T \sigma$ for an $n \times n$ matrix T with $T_{s,t} = x_s x_t \eta_{s,t}$.

Our main theorem follows by applying the following three lemmas. The first two lemmas give high probability bounds on the Frobenius and operator norms of T, and are proven in Section 5 and Section 6 respectively. The third lemma gives a central moment bound for quadratic forms in terms of both the Frobenius and operator norms of the associated matrix, and was proven in [11, 14][3]. By applying the bound of the third lemma to the quadratic form $\sigma^T T \sigma$, conditioned on the high probability event that both $\|T\|_F$ and $\|T\|_2$ are small, we obtain our main theorem.

Henceforth in this paper, we assume $\|x\|_2 = 1$, $\|x\|_\infty \le c$, and T is the matrix described above.

Lemma 4. $\mathbf{Pr}_h[\|T\|_F^2 > 4/k] < \delta$.

Lemma 5. $\mathbf{Pr}_h[\|T\|_2 > \varepsilon/(128 \cdot \log(1/\delta))] < \delta$.

Lemma 6 ([11, Theorem 5.1][4], [14])**.** *Let $z = (z_1, \ldots, z_n)$ be a vector of i.i.d. Bernoulli ± 1 random variables. Then for any $B \in \mathbb{R}^{n \times n}$ and even integer $\ell \ge 2$,*

$$\mathbf{E}\left[\left(z^T B z - \mathrm{trace}(B) \right)^\ell \right] \le 64^\ell \cdot \max\left\{ \sqrt{\ell} \cdot \|B\|_F, \ell \cdot \|B\|_2 \right\}^\ell.$$

[3] [14] proves a tail bound, but it is not hard to then derive a moment bound via integration; [11] directly proves a moment bound.

[4] What are denoted $\|B\|_F$ and $\|B\|_2$ here were denoted $\|B\|_2$ and $\|B\|_\infty$, respectively, in [11].

Theorem 7 (Main Theorem)**.**

$$\mathbf{Pr}_{h,\sigma}[|\|Ax\|_2^2 - 1| > \varepsilon] < 3\delta.$$

Proof. Write

$$\begin{aligned}\|Ax\|_2^2 &= \|x\|_2^2 + 2 \sum_{(s,t)\in\binom{[d]}{2}} x_s x_t \eta_{s,t}\sigma(s)\sigma(t) \\ &= 1 + Z.\end{aligned}$$

We will show $\mathbf{Pr}_{h,\sigma}[|Z| > \varepsilon] < 3\delta$. Condition on h, and let $\mathcal{E}$ be the event that $\|T\|_F^2 \le 4/k$ and $\|T\|_2 \le \varepsilon/\log(1/\delta)$. By applications of Lemma 4 and Lemma 5 and a union bound,

$$\mathbf{Pr}_{h,\sigma}[|Z| > \varepsilon] < \mathbf{Pr}_{\sigma}[|Z| > \varepsilon \mid \mathcal{E}] + 2\delta.$$

By a Markov bound applied to the random variable Z^ℓ for ℓ an even integer,

$$\mathbf{Pr}_{\sigma}[|Z| > \varepsilon \mid \mathcal{E}] < \mathbf{E}_{\sigma}[Z^\ell \mid \mathcal{E}]/\varepsilon^\ell.$$

Since $Z = \sigma^T T\sigma$ and $\mathrm{trace}(T) = 0$, applying Lemma 6 with $B = T$ and $\ell \le r_\sigma$ gives

$$\mathbf{Pr}_{\sigma}[|Z| > \varepsilon \mid \mathcal{E}] < 64^\ell \cdot \max\left\{\varepsilon^{-1}\sqrt{\frac{4\ell}{k}}, \frac{\ell}{128 \cdot \log(1/\delta)}\right\}^\ell. \tag{2}$$

since the ℓth moment is determined by r_σ-wise independence of σ. We conclude the proof by noting that the expression in Eq. (2) is at most δ for $\ell = \log(1/\delta)$. ■

5 A high probability bound on $\|T\|_F$

In this section we prove Lemma 4.

Proof (of Lemma 4). Recall that for $s, t \in [d]$, $\eta_{s,t}$ is the random variable indicating that $s \ne t$ and $h(s) = h(t)$. Then, Eq. (1) implies that $\|T\|_F^2 = \sum_{s<t} x_s^2 x_t^2 \eta_{s,t}$. Note $\|T\|_F^2$ is a random variable depending only on h. The plan of our proof is to directly bound the ℓth moment of $\|T\|_F^2$ for some large ℓ (specifically, $\ell = \Theta(\log(1/\delta))$), then conclude by applying Markov's inequality to the random variable $\|T\|_F^{2\ell}$. We bound the ℓth moment of $\|T\|_F^2$ via some combinatorics.

We now give the details of our proof. Consider the expansion $(\|T\|_F^2)^\ell$. We have

$$(\|T\|_F^2)^\ell = \sum_{\substack{s_1,\ldots,s_\ell \\ t_1,\ldots,t_\ell \\ \forall i\in[\ell] s_i<t_i}} \prod_{i=1}^{\ell} x_{s_i}^2 x_{t_i}^2 \eta_{s_i,t_i} \tag{3}$$

Let $\mathcal{G}_\ell$ be the set of all isomorphism classes of graphs (possibly containing multi-edges) with between 2 and 2ℓ unlabeled vertices, minimum degree at least 1, and exactly ℓ edges with distinct labels in $[\ell]$. We now define a map $f : \{\binom{[d]}{2}^\ell\} \to \mathcal{G}_\ell$; i.e. f maps the monomials in Eq. (3) to elements of $\mathcal{G}_\ell$. Focus on one monomial in Eq. (3) and let $S = \{s_1, \ldots, s_\ell, t_1, \ldots, t_\ell\}$. We map the monomial to an

$|S|$-vertex element of $\mathcal{G}_\ell$ as follows: associate each $u \in S$ with a vertex, and for each s_i, t_i, draw an edge from the vertices associated with s_i, t_i using edge label i.

We now analyze the expectation of the summation in Eq. (3) by grouping monomials which map to the same elements of $\mathcal{G}_\ell$ under f.

$$\mathbf{E}_h\left[(\|T\|_F^2)^\ell\right] = \sum_{G\in\mathcal{G}_\ell} \sum_{\substack{\{(s_i,t_i)\}\in\binom{[d]}{2}^\ell \\ f(\{(s_i,t_i)\})=G}} \left(\prod_{i=1}^{\ell} x_{s_i}^2 x_{t_i}^2\right)\cdot \mathbf{E}_h\left[\prod_{i=1}^{\ell}\eta_{s_i,t_i}\right]. \tag{4}$$

Observe that $\prod_{i=1}^{\ell}\eta_{s_i,t_i}$ is determined by $h(s_i), h(t_i)$ for each $i \in [\ell]$, and hence its expectation is determined by 2ℓ-wise independence of h. Note that this product is 1 if s_i and t_i hash to the same element for each i and is 0 otherwise. Each s_i, t_i pair hash to the same element if and only if for each connected component of G, all elements of $S = \{s_1, \ldots, s_\ell, t_1, \ldots, t_\ell\}$ corresponding to vertices in that component hash to the same value. For the v_G elements we are concerned with, where $v_G = |S|$ is the number of vertices in G, we can choose one element of $[k]$ for each connected component. Hence the number of possible values of h on S that cause $\prod_{i=1}^{\ell}\eta_{s_i,t_i}$ to be 1 is k^{m_G}, where G has m_G connected components. Each possibility happens with probability k^{-v_G}. Hence $\mathbf{E}_h[\prod_{i-1}^{\ell}\eta_{s_i,t_i}] = k^{m_G - v_G}$.

Also, consider the term $\prod_{i=1}^{\ell} x_{s_i}^2 x_{t_i}^2 = \prod_{i=1}^{v_G} x_{r_i}^{2\cdot\ell_i}$, where $S = \{r_i\}_{i=1}^{v_G}$, each ℓ_i is at least 1, and $\sum_i \ell_i = 2\ell$ (ℓ_i is just the degree of the vertex associated with r_i in G). Then,

$$\prod_{i=1}^{v_G} x_{r_i}^{2\cdot\ell_i} = \left(\prod_{i=1}^{v_G} x_{r_i}^{2\cdot(\ell_i-1)}\right)\cdot\left(\prod_{i=1}^{v_G} x_{r_i}^2\right) \le \left(\prod_{i=1}^{v_G} x_{r_i}^{2\cdot(\ell_i-1)}\right)\cdot\left(\prod_{i=1}^{v_G} x_{r_i}^2\right) \le c^{2(2\ell-v_G)}\cdot\left(\prod_{i=1}^{v_G} x_{r_i}^2\right).$$

Note then that the monomials $(\prod_{i=1}^{v_G} x_{r_i}^2)$ that arise from the summation over $\{(s_i,t_i)\} \in \binom{d}{2}^\ell$ with $f(\{(s_i,t_i)\}) = G$ in Eq. (4) are a subset of those monomials which appear in the expansion of $(\sum_{i=1}^{d} x_i^2)^{v_G} = 1$. Thus, plugging back into Eq. (4),

$$\mathbf{E}_h\left[(\|T\|_F^2)^\ell\right] \le \sum_{G\in\mathcal{G}_\ell} \frac{c^{2(2\ell-v_G)}}{k^{v_G-m_G}}. \tag{5}$$

Note the value ℓ in the $c^{2(2\ell-v_G)}$ term just arose as e_G, the number of edges in G. We bound the above summation by considering all ways to form an element of $\mathcal{G}_\ell$ by adding one edge at a time, starting from the empty graph G_0 with zero vertices and edges. In fact we will overcount some $G \in \mathcal{G}_\ell$, but this is acceptable since we only want an upper bound on Eq. (5).

Define $F(G) = c^{2(2e_G-v_G)}/k^{v_G-m_G}$. Initially we have $F(G_0) = 1$. We will add ℓ edges in order by label, from label 1 to ℓ. For the ith edge we have three options to form G_i from G_{i-1}: (a) we can add the edge between two existing vertices in G_{i-1}, (b) we can add two new vertices to G_{i-1} and place the edge between them, or (c) we can create one new vertex and connect it to an already-existing vertex of G_{i-1}. For each of these three options, we will argue that $n_i \cdot F(G_i)/F(G_{i-1}) \le 1/k$, where n_i is the number of ways to perform the operation we chose at step i. This implies that the summation in Eq. (5) is at most $(3/k)^\ell$ since at each step of forming an element of $\mathcal{G}_\ell$ we have three options for how to form G_i from G_{i-1}.

Let e be the number of edges, v the number of vertices, and m the number of connected components for some G_{i-1}. In option (a), v remains constant, e increases by 1, and m either remains constant or decreases by 1. In any case, $F(G_i)/F(G_{i-1}) \le c^4$, and $n_i < 2\ell^2$; the latter is because we have $\binom{v}{2} < 2\ell^2$ choices of vertices to connect. In option (b), $n_i = 1$, v increases by 2, e increases by 1, and m increases by 1, implying $n_i \cdot F(G_i)/F(G_{i-1}) = 1/k$. Finally, in option (c), $n_i = v \le 2\ell$, v increases by 1, e increases by 1, and m remains constant, implying $n_i \cdot F(G_i)/F(G_{i-1}) \le 2\ell c^2/k$. Thus, regardless of which of the three options we choose, $n_i \cdot F(G_i)/F(G_{i-1}) \le \max\{2\ell^2 c^4, 1/k, 2\ell c^2/k\}$, which is $1/k$ for $\ell = O(\log(1/\delta))$.

As discussed above, when combined with Eq. (5) this gives $\mathbf{E}_h[(\|T\|_F^2)^\ell] \le (3/k)^\ell$. Then, by Markov's inequality on the random variable $(\|T\|_F^2)^\ell$ for $\ell \ge 2$ and even, and assuming $2\ell \le r_h$,

$$\mathbf{Pr}_h[\|T\|_F^2 > 4/k] < (k/4)^\ell \cdot \mathbf{E}_h[(\|T\|_F^2)^\ell] < (3/4)^\ell,$$

which is at most δ for $\ell = \Theta(\log(1/\delta)) < r_h$. ■

6 A high probability bound on $\|T\|_2$

In this section we prove Lemma 5. For each $j \in [k]$ we use α_j to denote $\sum_{\substack{i\in[d]\\ h(i)=j}} x_i^2$.

Lemma 8. $\|T\|_2 \le \max\{c^2, \max_{j\in[k]} \alpha_j\}$.

Proof. Define the diagonal matrix R with $R_{i,i} = x_i^2$, and put $S = T + R$. For each $j \in [k]$, consider the vector v_j whose support is $h^{-1}(j)$, with $(v_j)_i = x_i$ for each i in its support. Then $S = \sum_{j=1}^k v_j \cdot v_j^T$. Thus $\text{rank}(S)$ is equal to the number of non-zero v_j, since they are clearly linearly independent (they have disjoint support and are thus orthogonal) and span the image of S. Furthermore, these non-zero v_j are eigenvectors of S since $Sv_j = \alpha_j v_j$, and are the only eigenvectors of S with non-zero eigenvalue since if u is perpendicular to all such v_j then $Au = 0$.

Now, $\|T\|_2 = \sup_{\|x\|_2=1} |x^T T x| = \sup_{\|x\|_2=1} |x^T S x - x^T R x|$. Since S, R are both positive semidefinite, we then have $\|T\|_2 \le \max\{\|S\|_2, \|R\|_2\}$. $\|R\|_2$ is clearly $\|x\|_\infty^2 \le c^2$, and we saw above that $\|S\|_2 = \max_{j\in[k]} \alpha_j$. ■

We will need the following form of the Chernoff bound in our proof of Lemma 5, as well as the following facts concerning the Gaussian and Gamma distributions.

Theorem 9 ([21, Theorem 2])**.** *Let $X_1, \ldots, X_n$ be independent scalar random variables with $|X_i| \le K$ almost surely, with mean μ_i and variance σ_i^2. Then for any $\lambda > 0$, one has*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^n X_i - \mu\right| \ge \lambda\sigma\right] \le C_1 \cdot \max\left\{\exp(-C_2\lambda^2), \exp(C_2\lambda\sigma/K)\right\}.$$

for some absolute constants $C_1, C_2 > 0$, where $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Fact 10. *The* Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ *with mean μ and variance σ^2 has density*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}},$$

and for $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\ell \ge 1$ an integer,

$$\mathbf{E}\left[|X-\mu|^\ell\right] = \sigma^\ell \cdot (\ell-1)!! \cdot \begin{cases} \sqrt{2/\pi}, & \text{for } \ell \text{ odd} \\ 1, & \text{for } \ell \text{ even} \end{cases}.$$

where $(2r)!! = r! \cdot 2^r$ and $(2r-1)!! = (2r)!/(r! \cdot 2^r)$ for $r \ge 1$.

Fact 11. *For $\ell, \theta > 0$, the function*

$$f(x) = x^{\ell-1} \cdot \frac{e^{-x/\theta}}{\theta^\ell \cdot \Gamma(\ell)}$$

is a probability density on $\mathbb{R}^+$ (the Gamma distribution $\Gamma(\ell, \theta)$*). In particular, $\int_0^\infty f(x) = 1$.*

We also make use of the following lemma in our proof of Lemma 5 in order to convert tail bounds for a probability distribution into moment bounds.

Lemma 12. *Let $\mathcal{D}$ be a distribution on $[0, \infty)$ with density function f and cumulative distribution function Φ. Let $\ell \geq 1$ be such that for $X \sim \mathcal{D}$, $\mathbf{E}[X^\ell]$ is finite and $\lim_{x\to\infty} x^\ell \cdot (1 - \Phi(x)) = 0$. Then*

$$\mathbf{E}[X^\ell] = \ell \cdot \int_0^\infty x^{\ell-1}(1 - \Phi(x))dx = \ell \cdot \int_0^\infty x^{\ell-1} \cdot \mathbf{Pr}_{X\sim\mathcal{D}}[X \geq x]dx.$$

Proof. Note $-f$ is the derivative of $1 - \Phi$ so that, by integration by parts,

$$\mathbf{E}[X^\ell] = -\left(\int_0^\infty x^\ell(-f(x)dx)\right) = -[x^\ell \cdot (1 - \Phi(x))]_0^\infty + \ell \cdot \int_0^\infty x^{\ell-1} \cdot (1 - \Phi(x))dx.$$

■

Proof (of Lemma 5). Our plan of the proof is as follows. First, we define β_j to be $\sum_{g(i)=j} x_i^2$ for a *truly random* hash function $g : [d] \to [k]$. Since we have full independence, we can apply Theorem 9 to obtain a tail bound for a single β_j. Integrating this tail bound then gives moment bounds. Since r_h-wise independence preserves ℓth moments for $\ell \leq r_h$, we can then use these moment bounds to argue tail bounds for the α_j under the r_h-wise independence of h. We conclude by applying a tail bound for a single α_j, then using a union bound over all $j \in [k]$.

Letting $\delta'_{i,j}$ indicate $g(i) = j$, we have $\beta_j = \sum_{i=1}^n x_i^2 \delta'_{i,j}$. Letting $X_i = x_i^2 \delta'_{i,j}$, in the notation of Theorem 9 we have $\mu = \|x\|_2^2/k = 1/k$ and $\sigma_i^2 \leq \mathbf{E}_h[X_i^2] = x_i^4/k$ so that $\sigma^2 = \sum_{i=1}^n x_i^4/k \leq \|x\|_\infty^2 \cdot \|x\|_2^2/k \leq c^2/k$. Furthermore, each X_i is never larger than c^2. Thus,

$$\mathbf{Pr}_h\left[\left|\beta_j - \frac{1}{k}\right| \geq \lambda\right] \leq C_1 \cdot \max\left\{e^{-C_2 \cdot \frac{\lambda^2 k}{c^2}}, e^{-C_2 \cdot \frac{\lambda}{c^2}}\right\}.$$

Applying Lemma 12 for even $\ell \geq 2$,

$$\begin{aligned}
\mathbf{E}_h\left[\left(\beta_j - \frac{1}{k}\right)^\ell\right] &\leq C_1\ell \cdot \int_0^\infty \lambda^{\ell-1} \cdot \max\left\{e^{-C_2 \cdot \frac{\lambda^2 k}{c^2}}, e^{-C_2 \cdot \frac{\lambda}{c^2}}\right\} d\lambda \\
&\leq C_1\ell \cdot \left(\int_0^{1/k} \lambda^{\ell-1} \cdot e^{-C_2 \cdot \frac{\lambda^2 k}{c^2}} d\lambda + \int_{1/k}^\infty \lambda^{\ell-1} \cdot e^{-C_2 \cdot \frac{\lambda}{c^2}} d\lambda\right) \\
&\leq C_1\ell \cdot \left(\int_{-\infty}^\infty |\lambda|^{\ell-1} \cdot e^{-C_2 \cdot \frac{\lambda^2 k}{c^2}} d\lambda + \int_0^\infty \lambda^{\ell-1} \cdot e^{-C_2 \cdot \frac{\lambda}{c^2}} d\lambda\right) \\
&= C_1\ell \cdot \left(\sqrt{2\pi} \cdot (\ell-2)!! \cdot \left(\frac{c^2}{2C_2 k}\right)^{\frac{\ell}{2}} + (\ell-1)! \cdot \left(\frac{c^2}{C_2}\right)^\ell\right) \qquad (6)
\end{aligned}$$

with the last equality using Fact 10 and Fact 11.

By Markov's inequality on the random variable $(\alpha_j - 1/k)^\ell$,

$$\mathbf{Pr}_h\left[\left|\alpha_j - \frac{1}{k}\right| > \lambda\right] < \frac{\mathbf{E}_h[(\alpha_j - 1/k)^\ell]}{\lambda^\ell}$$

for any even integer $\ell \geq 2$. As long as $\ell \leq r_h$ the moment bound of Eq. (6) holds for h by r_h-wise independence, in which case

$$\mathbf{Pr}_h\left[\left|\alpha_j - \frac{1}{k}\right| > \lambda\right] < 2^{O(\ell)} \cdot \left[\left(\frac{c\sqrt{\ell}}{\lambda\sqrt{k}}\right)^\ell + \left(\frac{c^2\ell}{\lambda}\right)^\ell\right] \tag{7}$$

by approximating the factorials of Eq. (6) via Stirling's formula (namely, $\ell! = \ell^\ell/2^{\Theta(\ell)}$). For $r_h = \Omega(\log(k/\delta))$, we can set $\ell = \Omega(\log(k/\delta))$ and $\lambda = \varepsilon/(256 \cdot \log(1/\delta))$ in Eq. (7) to obtain

$$\mathbf{Pr}_h\left[\left|\alpha_j - \frac{1}{k}\right| > \lambda\right] < \delta/k$$

as long as c is a sufficiently small constant times $\sqrt{\varepsilon/(\ell \cdot \log(1/\delta))}$. Then by a union bound over each j, we have $\mathbf{Pr}_h[\max_j \alpha_j > 1/k + \lambda > \varepsilon/(128 \cdot \log(1/\delta))] < \delta$. Our lemma then follows by applying Lemma 8, and using the fact that $c^2 < \varepsilon/(128 \cdot \log(1/\delta))$ for $\varepsilon < \varepsilon_0$. ■

Acknowledgments

This work was done while both authors were interns at Microsoft Research New England in Summer 2010. We thank Venkatesan Guruswami for showing us [21], which was very helpful.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th ACM Symposium on Theory of Computing (STOC)*, pages 557–563, 2006.

[3] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete Comput. Geom.*, 42(4):615–630, 2009.

[4] Nir Ailon and Edo Liberty. Almost optimal unrestricted fast Johnson-Lindenstrauss transform. *CoRR*, abs/1005.5513, 2010.

[5] Noga Alon. Problems and results in extremal combinatorics I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[6] Noga Alon and Joel H. Spencer. *The Probabilistic Method.* Wiley-Interscience, 2nd edition, 2000.

[7] Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[8] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 693–703, 2002.

[9] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

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

[11] Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. *CoRR*, abs/0911.3389, 2009.

[12] Lars Engebretsen, Piotr Indyk, and Ryan O'Donnell. Derandomized dimensionality reduction with applications. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 705–712, 2002.

[13] Peter Frankl and Hiroshi Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory. Ser. B*, 44(3):355–362, 1988.

[14] David Lee Hanson and F. Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 1971.

[15] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[16] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[17] Zohar Karnin, Yuval Rabani, and Amir Shpilka. Explicit dimension reduction and its applications. *Electronic Colloquium on Computational Complexity (ECCC)*, (121), 2009.

[18] Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and Lean Walsh transforms. In *Proceedings of the 12th International Workshop on Randomization and Computation (RANDOM)*, pages 512–522, 2008.

[19] Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC), to appear (see also CoRR abs/0910.4122)*, 2010.

[20] D. Sivakumar. Algorithmic derandomization via complexity theory. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–626, 2002.

[21] Terence Tao. Notes 1: Concentration of measure, 2010. `http://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/`.

[22] Mikkel Thorup and Yin Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 615–624, 2004.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de