Almost Optimal Explicit Johnson-Lindenstrauss Families

Daniel Kane[1], Raghu Meka[2], and Jelani Nelson[3]

[1] Harvard University, Cambridge, USA,
dankane@math.harvard.edu,
[2] University of Texas at Austin, USA,
raghu@cs.utexas.edu,
[3] MIT, Cambridge, USA,
minilek@mit.edu

Abstract. The Johnson-Lindenstrauss lemma is a fundamental result in probability with several applications in the design and analysis of algorithms. Constructions of linear embeddings satisfying the Johnson-Lindenstrauss property necessarily involve randomness and much attention has been given to obtain explicit constructions minimizing the number of random bits used. In this work we give explicit constructions with an almost optimal use of randomness: For $0 < \varepsilon, \delta < 1/2$, we obtain explicit generators $G : \{0,1\}^r \to \mathbb{R}^{s \times d}$ for $s = O(\log(1/\delta)/\varepsilon^2)$ such that for all d-dimensional vectors w of Euclidean norm 1,

$$\Pr_{y \in_u \{0,1\}^r}[\, |\|G(y)w\|^2 - 1| > \varepsilon \,] \leq \delta,$$

with seed-length $r = O\left(\log d + \log(1/\delta) \cdot \log\left(\frac{\log(1/\delta)}{\varepsilon}\right)\right)$. In particular, for $\delta = 1/\operatorname{poly}(d)$ and fixed $\varepsilon > 0$, we obtain seed-length $O((\log d)(\log\log d))$. Previous constructions required $\Omega(\log^2 d)$ random bits to obtain polynomially small error.

We also give a new elementary proof of the optimality of the JL lemma showing a lower bound of $\Omega(\log(1/\delta)/\varepsilon^2)$ on the embedding dimension. Previously, Jayram and Woodruff [9] used communication complexity techniques to show a similar bound.

1 Introduction

The celebrated Johnson-Lindenstrauss lemma (JLL) [10] is by now a standard technique for handling high dimensional data. Among its many known variants (see [4], [6], [8], [13]), we use the following version originally proven in [1], [4] [1].

Theorem 1. *For all $w \in \mathbb{R}^d$, $\|w\| = 1$, $0 < \varepsilon < 1/2$, $s \geq 1$,*

$$\Pr_{S \in_u \{1,-1\}^{s \times d}}[\, |\, \|(1/\sqrt{s})\, Sw\|^2 - 1 \,| \geq \varepsilon \,] \leq C \cdot e^{-C'\varepsilon^2 s}.$$

[1] Throughout, C, C' denote universal constants. For a multiset S, $x \in_u S$ denotes a uniformly random element of S. For $w \in \mathbb{R}^d$, $\|w\|$ denotes the Euclidean norm of w.

We say a family of random matrices has the JL property (or is a JL family) if the above condition holds. In typical applications of JLL, the error δ is taken to be $1/\operatorname{poly}(d)$ and the goal is to embed a given set of $\operatorname{poly}(d)$ points in d dimensions to $O(\log d)$ dimensions with distortion at most $1+\varepsilon$ for a fixed constant ε. This is the setting we concern ourselves with.

Linear embeddings of Euclidean space as above necessarily require randomness as else one can take the vector w to be in the kernel of the fixed transformation. To formalize this we use the following definition.

Definition 1. *For $\varepsilon, \delta > 0$, a generator $G : \{0,1\}^r \to \mathbb{R}^{s\times d}$ is a $(d, s, \delta, \varepsilon)$-JL generator of seed-length r if for every $w \in \mathbb{R}^d$, $\|w\| = 1$,*

$$\Pr_{y\in_u \{0,1\}^r}[\,|\,\|G(y)w\|^2 - 1\,| \geq \varepsilon\,] \leq \delta.$$

1.1 Derandomizing JLL

A simple probabilistic argument shows that there exists a $(d, O(\log(1/\delta)/\varepsilon^2), \delta, \varepsilon)$-JL generator with seed-length $r = O(\log d + \log(1/\delta))$. On the other hand, despite much attention the best known explicit generators have seed-length at least $\min(\,\Omega(\log(1/\delta)\log d),\ \Omega(\log d + \log^2(1/\delta))\,)$ [5], [11]. Besides being a natural problem in geometry as well as derandomization, an explicit JL generator with minimal randomness would likely help derandomize other geometric algorithms and metric embedding constructions. Further, having an explicit construction is of fundamental importance for streaming algorithms as storing the entire matrix (as opposed to the randomness required to generate the matrix) is often too expensive in the streaming context.

Our main result is an explicit generator that takes roughly $O((\log d)(\log\log d))$ random bits and outputs a matrix $A \in \mathbb{R}^{s\times d}$ satisfying the JL property for constant ε and $\delta = 1/\operatorname{poly}(d)$.

Theorem 2 (Main). *For every $0 < \varepsilon, \delta < 1/2$, there exists an explicit $(d, C\log(1/\delta)/\varepsilon^2, \delta, \varepsilon)$-JL generator $G : \{0,1\}^r \to \mathbb{R}^{s\times d}$ with seed-length*

$$r = O\left(\log d + \log(1/\delta)\cdot\log\left(\frac{\log(1/\delta)}{\varepsilon}\right)\right).$$

We give two different constructions. Our constructions are elementary in nature using only standard tools in derandomization such as k-wise independence and oblivious samplers [15]. Our first construction is simpler and gives a generic template for derandomizing most known JL families. The second construction has the advantage of allowing fast matrix-vector multiplications: the matrix-vector product $G(y)w$ can be computed efficiently in time $O(d\log d) + \operatorname{poly}(\log(1/\delta)/\varepsilon)$[2].

Further, as one of the motivations for derandomizing JLL is its potential applications in streaming, it is important that the entries of the generated matrices

[2] The computational efficiency does not follow directly from the dimensions of $G(y)$, as our construction involves composing matrices of much higher dimension.

be computable in small space. We observe that for any $i \in [s], j \in [d], y \in \{0,1\}^r$, the entry $G(y)_{ij}$ can be computed in space $O(\log d \cdot \mathrm{poly}(\log\log d))$ and time $O(d^{1+o(1)})$ (for fixed ε, $\delta > 1/\mathrm{poly}(d)$). (See proof of Theorem 8 for the exact bound.)

1.2 Optimality of JLL

We also give a new proof of the optimality of the JL lemma showing a lower-bound of $s_{\mathrm{opt}} = \Omega(\log(1/\delta)/\varepsilon^2)$ for the target dimension. Previously, Jayram and Woodruff [9] used communication complexity techniques to show a similar bound in the case $s_{\mathrm{opt}} < d^{1-\gamma}$ for some fixed constant $\gamma > 0$. In contrast, our argument is more direct in nature and is based on linear algebra and elementary properties of the uniform distribution on the sphere, and only requires the assumption $s_{\mathrm{opt}} < d/2$. Note the JLL is only interesting for $s_{\mathrm{opt}} < d$.

Theorem 3. *There exists a universal constant $c > 0$, such that for any distribution $\mathcal{A}$ over linear transformations from $\mathbb{R}^d$ to $\mathbb{R}^s$ with $s < d/2$, there exists a vector $w \in \mathbb{R}^d$, $\|w\| = 1$, such that* $\Pr_{S\sim\mathcal{A}}[\,|\|Sw\|^2 - 1| > \varepsilon\,] \geq \exp(-c(s\varepsilon^2 + 1))$.

1.3 Related Work

The ℓ_2 streaming sketch of Alon et al. [3] implies an explicit distribution over ℓ_2-embeddings with seed-length $O(\log d)$ for embedding $\mathbb{R}^d$ into $\mathbb{R}^s$ with distortion $1+\varepsilon$ and error δ, where $s = O(1/(\varepsilon^2\delta))$. Karnin et al. [11] construct an explicit JL family with optimal target dimension and seed-length $(1+o(1))\log d + O(\log^2(1/(\varepsilon\delta)))$. Clarkson and Woodruff [5] showed that a random scaled sign matrix with $O(\log(1/\delta))$-wise independent entries satisfies the JL lemma, giving seed-length $O(\log(1/\delta)\log d)$. We make use of their result in our construction.

We also note that there are efficient non-black box derandomizations of JLL, [7], [14]. These works take as input n points in $\mathbb{R}^d$, and deterministically compute an embedding (that depends on the input set) into $\mathbb{R}^{O(\log n)/\varepsilon^2}$ which preserves all pairwise distances between the given set of n points.

1.4 Outline of Constructions

For intuition, suppose that $\delta > 1/d^c$ is polynomially small and ε is a constant. Our constructions are based on a simple iterative scheme: We reduce the dimension from d to $\tilde{O}(\sqrt{d})$ (we say $f = \tilde{O}(g)$ if $f = O(g \cdot \mathrm{polylog}(g))$) and iterate for $O(\log\log d)$ steps.

Generic Construction. Our first construction gives a generic template for reducing the randomness required in standard JL families and is based on the following simple observation. Starting with any JL family, such as the random sign matrix construction of Theorem 1, there is a trade-off that we can make between the amount of independence required to generate the matrix and the final embedding dimension. For instance, if we only desire to embed to a dimension of $O(\sqrt{d})$ (as opposed to $O(\log d)$), it suffices for the entries of the random sign matrix to be $O(1)$-wise independent. We exploit this idea by iteratively decreasing the dimension from d to $O(\sqrt{d})$ and so on by using a random sign matrix with an increasing amount of independence at each iteration.

Fast JL Construction. Fix a vector $w \in \mathbb{R}^d$ with $\|w\| = 1$ and suppose $\delta = 1/\operatorname{poly}(d)$. We first use an idea of Ailon and Chazelle [2] who give a family of unitary transformations $\mathcal{R}$ from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that for every $w \in \mathbb{R}^d$ and $V \in_u \mathcal{R}$, the vector Vw is *regular*, in the sense that $\|Vw\|_\infty = O(\sqrt{(\log d)/d})$, with high probability. We derandomize their construction using limited independence to get a family of rotations $\mathcal{R}$ such that for $V \in_u \mathcal{R}$, $\|Vw\|_\infty = O(d^{-(1/2-\alpha)})$ with high probability, for a sufficiently small constant $\alpha > 0$.

We next observe that for a vector $w \in \mathbb{R}^d$, with $\|w\|_\infty = O(d^{-(1/2-\alpha)}\|w\|_2)$ projecting onto a random set of $O(d^{2\alpha}\log(1/\delta)/\varepsilon^2)$ coordinates preserves the ℓ_2 norm with distortion at most ε with high probability. We then note that the random set of coordinates can be chosen using oblivious samplers as in [15]. The idea of using samplers is due to Karnin et al. [11] who use samplers for a similar purpose.

Finally, iterating the above scheme $O(\log\log d)$ times we obtain an embedding of $\mathbb{R}^d$ to $\mathbb{R}^{\operatorname{poly}(\log d)}$ using $O(\log d \log\log d)$ random bits. We then apply the result of Clarkson and Woodruff [5] and perform the final embedding into $O(\log(1/\delta)/\varepsilon^2)$ dimensions by using a random scaled sign matrix with $O(\log(1/\delta))$-wise independent entries.

As all of the matrices involved in the construction are either Hadamard matrices or projection operators, the final embedding can actually be computed in $O(d\log d + \operatorname{poly}(\log(1/\delta)/\varepsilon))$ time.

Outline of Lowerbound. To show a lowerbound on the embedding dimension s, we use Yao's min-max principle to first transform the problem to that of finding a hard distribution on $\mathbb{R}^d$, such that no single linear transformation can embed a random vector drawn from the distribution well with very high probability. We then show that the uniform distribution over the d-dimensional sphere is one such hard distribution. The proof of the last fact involves elementary linear algebra and some direct calculations.

2 Preliminaries

We first state the classical Khintchine-Kahane inequalities (cf. [12]) which give tight moment bounds for linear forms.

Lemma 1 (Khintchine-Kahane). *For every $w \in \mathbb{R}^n$, $x \in_u \{1,-1\}^n$, $k > 0$,*

$$\mathbb{E}[\,|\langle w, x\rangle|^k\,] \le k^{k/2}\,\mathbb{E}[\,|\langle w, x\rangle|^2\,]^{k/2} = k^{k/2}\|w\|^k.$$

We use randomness efficient oblivious samplers due to Zuckerman [15] (See Theorem 3.17 and the remark following the theorem in [15]).

Theorem 4 (Zuckerman [15]). *There exists a constant C such that for every $\varepsilon, \delta > 0$ there exists an explicit collection of subsets of $[d]$, $\mathcal{S}(d,\varepsilon,\delta)$, with each $S \in \mathcal{S}$ of cardinality $|S| = s(\varepsilon,\delta,d) = ((\log d + \log(1/\delta))/\varepsilon)^C$, such that for every function $f : [d] \to [0,1]$,*

$$\Pr_{S \in_u \mathcal{S}}\left[\left|\frac{1}{s}\sum_{i\in S} f(i) - \mathop{\mathbb{E}}_{i \in_u [d]} f(i)\right| > \varepsilon\right] \le \delta,$$

and there exists an NC algorithm that generates random elements of $\mathcal{S}$ using $O(\log d + \log(1/\delta))$ random bits.

Corollary 1. *There exists a constant C such that the for every $\varepsilon, \delta, B > 0$ there exists an explicit collection of subsets of $[d]$, $\mathcal{S}(d, B, \varepsilon, \delta)$, with each $S \in \mathcal{S}$ of cardinality $|S| = s(d, B, \varepsilon, \delta) = ((\log d + \log(1/\delta))B/\varepsilon)^C$, such that for every function $f : [d] \to [0, B]$,*

$$\Pr_{S \in_u \mathcal{S}} \left[\left| \frac{1}{s} \sum_{i \in S} f(i) - \mathop{\mathbb{E}}_{i \in_u [d]} f(i) \right| > \varepsilon \right] \leq \delta,$$

and there exists an NC algorithm that generates random elements of $\mathcal{S}$ using $O(\log d + \log(1/\delta))$ random bits.

Proof. Apply the above theorem to $\bar{f} : [d] \to [0, 1]$ defined by $\bar{f}(i) = f(i)/B$.

Let $H_d \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^{d \times d}$ be the normalized Hadamard matrix such that $H_d^T H_d = I_d$ (we drop the suffix d when dimension is clear from context). While the Hadamard matrix is known to exist for powers of 2, for clarity, we ignore this technicality and assume that it exists for all d. Finally, let $\mathcal{S}^{d-1}$ denote the Euclidean sphere $\{w : w \in \mathbb{R}^d, \|w\| = 1\}$.

The following definitions will be useful in giving an abstract description of our constructions.

Definition 2. *A distribution $\mathcal{D}$ over $\mathbb{R}^{s \times d}$ is said to be a $(d, s, \delta, \varepsilon)$-*JL distribution *if for any $w \in \mathcal{S}^{d-1}$, $\Pr_{S \sim \mathcal{D}} \left[\left| \|Sw\|^2 - 1 \right| > \varepsilon \right] < \delta$.*

Definition 3. *A distribution $\mathcal{D}$ over $\mathbb{R}^{s \times d}$ is said to have the $(d, s, t, \delta, \varepsilon)$-JL moment property if for any $w \in \mathcal{S}^{d-1}$, $\mathbf{E}_{S \sim \mathcal{D}} \left[\left| \|Sw\|^2 - 1 \right|^t \right] < \varepsilon^t \cdot \delta$.*

Definition 4. *A distribution $\mathcal{D}$ is called a* strong (d, s)-JL distribution *if it is a $(d, s, \exp(-\Omega(\min\{\varepsilon, \varepsilon^2\} \cdot s)), \varepsilon)$-JL distribution for all $\varepsilon > 0$. If $\mathcal{D}$ has the $(d, s, \ell, O(\max\{\sqrt{\ell/(\varepsilon^2 s)}, \ell/(\varepsilon s)\})^\ell, \varepsilon)$-JL moment property for all $\varepsilon > 0$ and integer $\ell \geq 2$, then we say $\mathcal{D}$ has the* strong (d, s)-JL moment property.

Theorem 1 shows the conditions for being a strong (d, s)-JL distribution are met by random Bernoulli matrices when $0 < \varepsilon \leq 1$, though in fact the conditions are also met for all $\varepsilon > 0$ (see the proof in [5] for example). Sometimes we omit the d, s terms in the notation above if these quantities are clear from context, or if it is not important to specify them.

Throughout, we let logarithms be base-2 and often assume various quantities, like $1/\varepsilon$ or $1/\delta$, are powers of 2; this is without loss of generality.

3 Strong JL Distributions

It is not hard to show that having the strong JL moment property and being a strong JL distribution are equivalent. We use the following standard fact.

Fact 5 *Let Y, Z be nonnegative random variables such that* $\Pr[Z \geq t] = O(\Pr[Y \geq t])$ *for any* $t \geq 0$*. Then for* $\ell \geq 1$ *if* $\mathbf{E}[Y^\ell] < \infty$*, we have* $\mathbf{E}[Z^\ell] = O(\mathbf{E}[Y^\ell])$*.*

Theorem 6. *A distribution $\mathcal{D}$ is a strong (d, s)-JL distribution if and only if it has the strong (d, s)-JL moment property.*

Proof. First assume $\mathcal{D}$ has the strong JL moment property. Then, for arbitrary $w \in \mathcal{S}^{d-1}, \varepsilon > 0$,

$$\Pr_{S\sim\mathcal{D}}[|\|Sw\|^2 - 1| > \varepsilon] < \varepsilon^{-\ell} \cdot \mathbf{E}[|\|Sw\|^2 - 1|^\ell] < O(\max\{\sqrt{\ell/(\varepsilon^2 s)}, \ell/(\varepsilon s)\})^\ell.$$

The claim follows by setting $\ell = O(\min\{\varepsilon, \varepsilon^2\} \cdot s)$.

Now assume $\mathcal{D}$ is a strong JL distribution. Set $Z = |\|Sw\|^2 - 1|$. Since $\mathcal{D}$ is a strong JL distribution, the right tail of Z is big-Oh of that of the absolute value of the nonnegative random variable Y which is the sum of a Gaussian with mean 0 and variance $O(1/s)$, and an exponential random variable with parameter s. Now, apply Fact 5.

Remark 1. Theorem 6 implies that any strong JL distribution can be derandomized using $2\log(1/\delta)$-wise independence giving an alternate proof of the derandomized JL result of Clarkson and Woodruff (Theorem 2.2 in [5]). This is because, by Markov's inequality with ℓ even, and for $\varepsilon < 1$,

$$\Pr_{S\sim\mathcal{D}}\left[\left|\|Sw\|^2 - 1\right| > \varepsilon\right] < \varepsilon^{-\ell} \cdot \mathbf{E}_{S\sim\mathcal{D}}\left[\left(\|Sw\|^2 - 1\right)^\ell\right] \leq 2^{O(\ell)} \cdot (\varepsilon^{-1} \cdot \sqrt{\ell/s})^\ell. \tag{3.1}$$

Setting $\ell = \log(1/\delta)$ and $s = C\ell/\varepsilon^2$ for $C > 0$ sufficiently large makes the above probability at most δ. Now, note the ℓth moment is determined by 2ℓ-wise independence of the entries of S.

4 A Generic JL Derandomization Template

Theorem 6 and Remark 1 provide the key insight for our construction. If we use $\ell = 2\log(1/\delta)$-wise independent Bernoulli entries as suggested in Remark 1, the seed length would be $O(\ell \log d) = O(\log(1/\delta)\log d)$ for $s = \Theta(\varepsilon^{-2}\log(1/\delta))$. However, note that in Eq. (3.1), a trade-off can be made between the amount of independence needed and the final embedding dimension without changing the error probability. In particular, it suffices to use 4-wise independence if we embed into $s = \Omega(\varepsilon^{-2}\delta^{-1})$ dimensions. In general, if $s = C\varepsilon^{-2}q$ for $\log^2(1/\delta) \leq q \leq 1/\delta$, it suffices to set $\ell = O(\log_q(1/\delta))$ to make the right hand side of Eq. (3.1) at most δ. By gradually reducing the dimension over the course of several iterations, using higher independence in each iteration, we obtain shorter seed length.

Our main construction is described in Figure 1. We first embed into $O(\varepsilon^{-2}\delta^{-1})$ dimension using 4-wise independence. We then iteratively project from $O(\varepsilon^{-2}\delta^{-1/2^i})$ dimensions into $O(\varepsilon^{-2}\delta^{-1/2^{i+1}})$ dimensions until we have finally embedded into $O(\varepsilon^{-2}\log^2(1/\delta))$ dimensions. In our final step, we embed into the optimal target dimension using $2\log(1/\delta)$-wise independence. Note the Bernoulli distribution is not special here; we could use any family of strong JL distributions.

Iterative dimensionality reduction:

```
// Output S distributed according to a (d, s, δ, ε)-JL
   distribution.
```

1. Define $m = \log((\log 1/\delta)/(2 \log\log 1/\delta))$, $\varepsilon' = \varepsilon/(e(m+2))$, $\delta' = \delta/(m+2)$.
2. Define $s_i = C(\varepsilon')^{-2}\delta'^{-1/2^i}$, $\ell_i = \Theta(2^i)$ an even integer for $i \geq 0$. Define $s_{-1} = d$.
3. Let S_i be a random matrix drawn from a distribution with the $(s_{i-1}, s_i, \ell_i, \delta', \varepsilon')$-JL moment property for $i = 0, \ldots, m$.
4. Let S_{final} be drawn from a $(s_m, O(\varepsilon^{-2}\log(1/\delta)), \delta', \varepsilon')$-JL distribution.
5. $S \leftarrow S_{\text{final}} \cdot S_m \cdots S_0$.

Fig. 1. A general derandomization scheme for distributions with JL moment properties.

Theorem 7. *The output matrix S in Figure 1 is distributed according to a $(d, s, \delta, \varepsilon)$-JL distribution for $s = O(\log(1/\delta)/\varepsilon^2)$.*

Proof. For a fixed vector w, let $w^i = S_i \cdots S_0 w$, and let w^{-1} denote w. Then by our choice of s_i and a Markov bound on the ℓ_ith moment,

$$\Pr\left[\|w^i\|^2 - \|w^{i-1}\|^2 \,|| > \varepsilon'\|w^{i-1}\|^2\right] < \varepsilon'^{-\ell_i} \cdot \mathbf{E}[(\|w^i\|^2/\|w^{i-1}\|^2 - 1)^{\ell_i}] < \delta'$$

for $0 \leq i \leq m$. We also have $\Pr\left[\|S_{\text{final}}w^m\|^2 - \|w^m\|^2 \,|| > \varepsilon'\|w^m\|^2\right] < \delta'$. By a union bound, $\|S_{\text{final}}w^m\|^2 \leq (1+\varepsilon')^{m+2} \leq e^{(m+2)\varepsilon'} \leq 1+\varepsilon$ with probability $1-(m+2)\delta' = 1-\delta$.

As a corollary, we obtain our main theorem, Theorem 2.

Proof (of Theorem 2). We let the distributions in Steps 3 and 4 of Figure 1 be strong JL distributions. Then Steps 3 and 4 are satisfied by Remark 1.[

The seed length required to generate S_0 is $O(\log d)$. For S_i for $i > 0$ the seed length is $O(\ell_i \log(\varepsilon'^{-2}\delta'^{1/2^i})) = O(2^i \log(1/\varepsilon') + \log(1/\delta'))$, which is never larger than $O((\log(1/\delta')/\log\log(1/\delta'))\log(1/\varepsilon') + \log(1/\delta'))$, which is $O((\log(1/\delta)/\log\log(1/\delta))\log(1/\varepsilon) + \log(1/\delta))$. The seed length required for S_{final} is $O(\log(1/\delta')\log(\log(1/\delta')/\varepsilon')) = O(\log(1/\delta)\log(\log(1/\delta)/\varepsilon))$. Thus, the total seed length is dominated by generating S_0 and S_{final}, giving the claim. The distorion and error probabilities can be bounded by a union bound.

5 Explicit JL Families via Samplers

We now give an alternate construction of an explicit JL family. The construction is similar in spirit to that of the previous section and has the additional property that matrix-vector products for matrices output by the generator can be computed in time roughly $O(d \log d + s^3)$, as it is based on the Fast Johnson-Lindenstrauss Transform (FJLT) of [2]. For clarity, we concentrate on the case of $\delta = \Theta(1/d^c)$ polynomially small. The case of general δ can be handled similarly with some minor technical issues[3] that we skip in this extended abstract. Further, we assume that $\log(1/\delta)/\varepsilon^2 < d$ as else JLL is not interesting.

As outlined in the introduction, we first give a family of rotations to regularize vectors in $\mathbb{R}^d$. For a vector $x \in \mathbb{R}^d$, let $D(x) \in \mathbb{R}^{d\times d}$ be the diagonal matrix with $D(x)_{ii} = x_i$.

Lemma 2. *Let $x \in \{1, -1\}^d$ be drawn from a k-wise independent distribution. Then, for every $w \in \mathbb{R}^d$ with $\|w\| = 1$, $0 < \alpha < 1/2$,*

$$\Pr[\, \|HD(x)w\|_\infty > n^{-(1/2-\alpha)} \,] \le \frac{k^{k/2}}{n^{\alpha k-1}}.$$

Proof. Let $v = HD(x)w$. Then, for $i \in [d]$, $v_i = \sum_j H_{ij}x_jw_j$ and $\mathbb{E}[v_i^2] = \sum_j H_{ij}^2 w_j^2 = 1/d$. By Markov's inequality and the Khintchine-Kahane inequality (Lemma 1),

$$\Pr[\, |v_i| > d^{-(1/2-\alpha)} \,] \le \mathbb{E}[v_i^k] \cdot d^{(1/2-\alpha)k} \le k^{k/2} d^{(1/2-\alpha)k}/d^{k/2} = k^{k/2} d^{-\alpha k}.$$

The claim now follows from a union bound over $i \in [d]$.

We now give a family of transformations for reducing d dimensions to $\tilde{O}(d^{1/2})\cdot$ poly(s_{opt}) dimensions using oblivious samplers. For $S \subseteq [d]$, let $\mathcal{P}_S : \mathbb{R}^d \to \mathbb{R}^{|S|}$ be the projection onto the coordinates in S. In the following let C be the universal constant from Corollary 1.

Lemma 3. *Let $\mathcal{S} \equiv \mathcal{S}(d, d^{1/2C}, \varepsilon, \delta)$, $s = O(d^{1/2}\log^C(1/\delta)/\varepsilon^C)$ be as in Corollary 1 and let $\mathcal{D}$ be a k-wise independent distribution over $\{1, -1\}^d$. For $S \in_u \mathcal{S}$, $x \leftarrow \mathcal{D}$, define the random linear transformation $A_{S,x} : \mathbb{R}^d \to \mathbb{R}^s$ by $A_{S,x} = \sqrt{d/s} \cdot \mathcal{P}_S \cdot HD(x)$. Then, for every $w \in \mathbb{R}^d$ with $\|w\| = 1$,*

$$\Pr[\, |\|A_{S,x}(w)\|^2 - 1| \ge \varepsilon \,] \le \delta + k^{k/2}/d^{k/4C-1}.$$

Proof. Let $v = HD(x)w$. Then, $\|v\| = 1$ and by Lemma 2 applied for $\alpha = 1/4C$,

$$\Pr[\, \|v\|_\infty > d^{-(1/2-1/4C)} \,] \le k^{k/2}/d^{k/4C-1}.$$

Now condition on the event $\|v\|_\infty \le d^{-(1/2-1/4C)}$. Define $f : [d] \to \mathbb{R}$ by $f(i) = d \cdot v_i^2 \le d^{1/2C} = B$. Then,

$$\|A_{S,x}(w)\|^2 = (d/s)\|\mathcal{P}_S(v)\|^2 = \frac{1}{s}\sum_{i\in S} dv_i^2 = \frac{1}{s}\sum_{i\in S} f(i),$$

[3] In case of very small δ, we need to ensure that we never increase the dimension - which can be done trivially by using the identity transformation. In case of large δ, we first embed the input vector into $O(1/\delta\varepsilon^2)$ dimensions using 4-wise independence as in Section 4.

and $\mathbb{E}_{i \in_u [d]} f(i) = (1/d) \sum_i d \cdot v_i^2 = 1$. Therefore, by Corollary 1,

$$\Pr[\,|\,\|A_{S,x}(w)\|^2 - 1\,| \geq \varepsilon\,] = \Pr_{S \in_u \mathcal{S}} \left[\left| \frac{1}{s} \sum_{i \in S} f(i) - \mathop{\mathbb{E}}_{i \in_u [d]} f(i) \right| \geq \varepsilon \right] \leq \delta.$$

The claim now follows.

We now recursively apply the above lemma. Fix $\varepsilon, \delta > 0$. Let $\mathcal{A}(d,k) : \mathbb{R}^d \to \mathbb{R}^{s(d)}$ be the collection of transformations $\{A_{S,x} : S \in_u \mathcal{S}, x \leftarrow D\}$ as in the above lemma for $s(d) = s(d, d^{1/2C}, \varepsilon, \delta) = c_1 d^{1/2} (\log d/\varepsilon)^C$, for a constant c_1. Note that we can sample from $\mathcal{A}(d,k)$ using $r(d,k) = k \log d + O(\log d + \log(1/\delta)) = O(k \log d)$ random bits.

Let $d_0 = d$, and let $d_{i+1} = s(d_i)$. Let $k_0 = 8C(c+1)$ (recall that $\delta = 1/d^c$) and $k_{i+1} = 2^i k_0$. The parameters d_i, k_i are chosen so that $1/d_i^{k_i}$ is always polynomially small. Fix $t > 0$ to be chosen later so that $k_i < d_i^{1/4C}$ for $i < t$.

Lemma 4. *For $A_0 \in_u \mathcal{A}(d_0, k_0), A_1 \in_u \mathcal{A}(d_1, k_1), \cdots, A_{t-1} \in_u \mathcal{A}(d_{t-1}, k_{t-1})$ chosen independently, and $w \in \mathbb{R}^d$, $\|w\| = 1$,*

$$\Pr[\,(1-\varepsilon)^t \leq \|A_{t-1} \cdots A_1 A_0(w)\|^2 \leq (1+\varepsilon)^t\,] \geq 1 - t\delta - \sum_{i=0}^{t-1} \frac{k_i^{k_i/2}}{d_i^{k_i/4C-1}}.$$

Proof. The proof is by induction on $i = 1, \ldots, t$. For $i = 1$, the claim is same as Lemma 3. Suppose the statement is true for $i-1$ and let $v = A_{i-1} \cdots A_0(w)$. Then, $v \in \mathbb{R}^{d_i}$ and the lemma follows by Lemma 3 applied to $\mathcal{A}(d_i, k_i)$, and v.

What follows is a series of elementary calculations to bound the seed-length and error from the above lemma. Observe that

$$d^{(1/2)^i} \leq d_i = d^{(1/2)^i} \cdot \left(\frac{c_1 \log^C(d)}{\varepsilon^C} \right)^{1+(1/2)+\cdots+(1/2)^{i-1}} \leq d^{(1/2)^i} \left(\frac{c_1 \log^C d}{\varepsilon^C} \right)^2. \tag{5.1}$$

Let $t = O(\log\log d)$ be such that $2^t = \log d / 4C \log\log d$. Then, $d_t \leq \log^{4C} d \cdot (c_1 \log^C d/\varepsilon^C)^2 = O(\log^{6C} d/\varepsilon^{2C})$, and for $i < t$,

$$k_i < k_t = 8C(c+1)2^t = 2(c+1)\log d/\log\log d < \log d = d^{(1/2)^t/4C} < d_t^{1/4C} < d_i^{1/4C}, \tag{5.2}$$

where we assumed that $\log\log d > 2c+2$. Therefore, the error in Lemma 4 can be bounded by

$$\begin{aligned}
t\delta + \sum_{i=0}^{t-1} \frac{k_i^{k_i/2}}{d_i^{k_i/4C-1}} &\leq t\delta + d \sum_{i=0}^{t-1} d_i^{-k_i/8C} && \text{(Equation 5.2)} \\
&\leq t\delta + d \sum_{i=0}^{t-1} (d^{1/2^i})^{-8C(c+1)\cdot 2^i/8C} && \text{(Equation 5.1)} \\
&\leq t\delta + t/d^c \leq 2t\delta && \text{(as } \delta > 1/d^c\text{)}.
\end{aligned}$$

Note that,

$$k_i \log d_i \le 8C(c+1) \cdot 2^i (\log d/2^i + 2C \log\log d + 2C \log(1/\varepsilon)) = O(\log d + \log d \log(1/\varepsilon)/\log\log d).$$

Therefore, the randomness needed after $t = O(\log\log d)$ iterations is

$$\sum_{i=0}^{t-1} O(k_i \log d_i) = O(\log d \log\log d + (\log d)\log(1/\varepsilon)).$$

Combining the above arguments (applied to $\delta' = \delta/\log\log d$ and $\varepsilon' = \varepsilon/\log\log d$ and simplifying the resulting expression for seed-length) we obtain our fast derandomized JL family.

Theorem 8 (Fast Explicit JL Family). *There exists a $(d, O(\log(1/\delta)/\varepsilon^2), \delta, \varepsilon)$-JL generator with seed-length $r = O(\log d + \log(1/\delta)(\log(\log(1/\delta)/\varepsilon)))$ such that for every vector $w \in \mathbb{R}^d$, $y \in \{0,1\}^r$, $G(y)w$ can be evaluated in time $O(d\log d + \mathrm{poly}(\log(1/\delta)/\varepsilon))$.*

Proof. We suppose that $\delta = \Theta(1/d^c)$ - the analysis for the general case is similar.

From the above arguments there is an explicit generator that takes $O(\log(d/\delta)\cdot \log(\,\log(d/\delta)/\varepsilon\,))$ random bits and outputs a linear transformation $A : \mathbb{R}^d \to \mathbb{R}^m$ for $m = \mathrm{poly}(\log(d/\delta), 1/\varepsilon)$, satisfying the JL property with error at most δ and distortion at most ε. The theorem now follows by composing the transformations of the above theorem with a sign matrix having $2\log(1/\delta)$-wise independent entries. The additional randomness required is $O(\log(1/\delta)\log m) = O(\log(1/\delta)(\log\log(d/\delta) + \log(1/\varepsilon))$.

We next bound the time for computing matrix-vector products for the matrices we output. Note that for $i < t$, the matrices A_i of Lemma 4 are of the form $\mathcal{P}_S \cdot H_{d_i} D(x)$ for a k-wise independent string $x \in \{1,-1\}^{d_i}$. Thus, for any vector $w_i \in \mathbb{R}^{d_i}$, $A_i w_i$ can be computed in time $O(d_i \log d_i)$ using the discrete Fourier transform. Therefore, for any $w = w_0 \in \mathbb{R}^{n_0}$, the product $A_{t-1}\cdots A_1 A_0 w_0$ can be computed in time

$$\sum_{i=0}^{t-1} O(d_i \log d_i) \le O(d\log d) + \log d \cdot \sum_{i=1}^{t-1} O\left(d^{1/2^i}(\log(1/\delta)/\varepsilon^2)^2\right) \qquad \text{(Equation 5.1)}$$
$$= O(d\log d + \sqrt{d}\log d \log^2(1/\delta)/\varepsilon^4).$$

The above bound dominates the time required to perform the final embedding.

A similar calculation shows that for indices $i \in s, j \in [d]$, the entry $G(y)_{ij}$ of the generated matrix can be computed in space $O\left(\sum_i \log d_i\right) = O(\log d + \log(1/\varepsilon)\cdot\log\log d)$ by expanding the product of matrices and enumerating over all intermediary indices[4]. The time required to perform the calculation is $O(s \cdot d_t \cdot d_{t-1} \cdots d_1) = d \cdot (\log d/\varepsilon)^{O(\log\log d)}$.

[4] We also need to account for the time and space needed by the samplers and for generating k-wise independent strings. However, these are dominated by the task of enumerating over all indices; for instance, the samplers of [15] are in NC.

Remark 2. We can use the FJLT of [2] in the framework of Figure 1 to get seed length and update time as above. Details are deferred to the full version.

6 Optimality of JL Lemma

We next prove Theorem 3, the optimality of the number of rows in the JL Lemma. Let $\mathcal{A}$ be a distribution over linear transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$ such that $\Pr_{S\sim\mathcal{A}}[|\|Sw\|^2 - 1|] < \delta]$ for any $w \in \mathcal{S}^{d-1}$. Then, it must be the case that $\Pr_{S\sim\mathcal{A}}[\Pr_{w\in_u\mathcal{S}^{d-1}}[|\|Sw\|^2-1|]] < \delta$. By an averaging argument, there must exist a linear transformation S in the support of $\mathcal{A}$ such that $\Pr_{w\in\mathcal{S}^{d-1}}[|\|Sw\|^2-1|] < \delta$. We show this cannot happen unless k is sufficiently large.

Theorem 9. *If $S : \mathbb{R}^d \to \mathbb{R}^k$ is a linear transformation with $d > 2k$ and $\varepsilon > 0$ sufficiently small, then for w a randomly chosen vector in $\mathcal{S}^{d-1}$, $\Pr[|\|Sw\|^2-1| > \varepsilon] \geq \exp(-O(k\varepsilon^2 + 1))$.*

Proof. First note that we can assume S is surjective as else, we may replace $\mathbb{R}^k$ by the image of S. Let $V = \ker(S)$ and let U be the orthogonal complement of V in $\mathbb{R}^d$. Then $\dim(U) = k$, $\dim(V) = d-k$. Now, any $w \in \mathbb{R}^d$ can be written uniquely as $w_V + w_u$ where w_V and w_u are the components of w in V and U respectively. We may then write $w_V = r_V \Omega_V$, $w_u = r_u\Omega_u$, where r_V, r_u are positive real numbers and Ω_V and Ω_u are unit vectors in V and U respectively.

Let $s_V = r_V^2$ and $s_u = r_u^2$. We may now parameterize the unit sphere by $(s_V, \Omega_V, s_u, \Omega_u) \in [0,1] \times \mathcal{S}^{d-k-1} \times [0,1] \times \mathcal{S}^{k-1}$, so that $s_V + s_u = 1$. It is clear that the uniform measure on the sphere is given in these coordinates by $f(s_u)ds_u d\Omega_V d\Omega_u$ for some function $f : [0,1] \to [0,1]$. We next show that

$$f(s_u) = C_f \cdot (1-s_u)^{(d-k-2)/2} s_u^{(k-2)/2}, \tag{6.1}$$

where C_f is a normalization constant. Observe that $f(s_u)$ should be proportional to the limit as $\delta_1, \delta_2 \to 0^+$ of $(\delta_1\delta_2)^{-1}$ times the volume of points w satisfying $\|w_u\|^2 \in [s_u, s_u + \delta_2]$ and $\|w_V\|^2 \in [1 - \|w_u\|^2, 1 - \|w_u\|^2 + \delta_1]$. For fixed w_u, the latter volume is within $O(\delta_1\delta_2)$ of the volume of w_V so that $\|w_V\|^2 \in [s_V, s_V + \delta_1]$. Now the measure on V is $r_V^{d-k-1} dr_V d\Omega_V$. Therefore it also is $\frac{1}{2}s_V^{(d-k-2)/2} ds_V d\Omega_V$. Therefore this volume over V is proportional to $s_V^{(d-k-2)/2}(\delta_1 + O(\delta_1\delta_2 + \delta_1^2))$. Similarly the volume of w_u so that $\|w_u\|^2 \in [s_u, s_u+\delta_2]$ is proportional to $s_u^{(k-2)/2}(\delta_2 + O(\delta_2^2))$. Hence f is proportional to $s_V^{(d-k-2)/2} s_u^{(k-2)/2}$.

We are now prepared to prove the theorem. The basic idea is to first condition on Ω_V, Ω_u. We let $C = \|S\Omega_u\|^2$. Then if w is parameterized by $(s_V, \Omega_V, s_u, \Omega_u)$, $\|Sw\|^2 = Cs_u$. Choosing w randomly, we know that $s = s_u$ satisfies the distribution $\frac{s^{(k-2)/2}(1-s)^{(d-k-2)/2}}{\beta((k-2)/2,(d-k-2)/2)}ds = f(s)ds$ on $[0,1]$. We need to show that for any $c = \frac{1}{C}$, the probability that s is not in $[(1-\varepsilon)c, (1+\varepsilon)c]$ is $\exp(-O(\varepsilon^2 k))$. Note that $f(s)$ attains its maximum value at $s_0 = \frac{k-2}{d-4} < \frac{1}{2}$. Notice that $\log(f(s_0(1+x)))$ is some constant plus $\frac{k-2}{2}\log(s_0(1+x)) + \frac{d-k-2}{2}\log(1-s_0-xs_0)$. If $|x| <$

$1/2$, then this is some constant plus $-O(kx^2)$. So for such w, $f(s_0(1+x)) = f(s_0)\exp(-O(kx^2))$. Furthermore, for all x, $f(s_0(1+x)) = f(s_0)\exp(-\Omega(kx^2))$. This says that f is bounded above by a normal distribution and checking the normalization we find that $f(s_0) = \Omega(s_0^{-1}k^{1/2})$.

We now show that both $\Pr(s < (1-\varepsilon)s_0)$ and $\Pr(s > (1+\varepsilon)s_0)$ are reasonably large. We can lower bound either as

$$\begin{aligned} s_0 \int_{\varepsilon}^{1/2} f(s_0)\exp(-O(kx^2))dx &\geq \Omega(k^{1/2}) \int_{\varepsilon}^{\varepsilon+k^{-1/2}} \exp(-O(kx^2))dx \\ &\geq \Omega(\exp(-O(k(\varepsilon+k^{-1/2})^2))) \\ &\geq \exp(-O(k\varepsilon^2+1)). \end{aligned}$$

Hence since one of these intervals is disjoint from $[(1-\varepsilon)c, (1+\varepsilon)c]$, the probability that s is not in $[(1-\varepsilon)c, (1+\varepsilon)c]$ is at least $\exp(-O(k\varepsilon^2+1))$.

References

1. D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.
2. N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.
3. N. Alon, Y. Matias, and M. Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.
4. R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.
5. K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.
6. S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.
7. L. Engebretsen, P. Indyk, and R. O'Donnell. Derandomized dimensionality reduction with applications. In *SODA*, pages 705–712, 2002.
8. P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC*, pages 604–613, 1998.
9. T. S. Jayram and D. P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with low error. In *SODA*, pages 1–10, 2011.
10. W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp Math*, 26:189–206, 1984.
11. Z. Karnin, Y. Rabani, and A. Shpilka. Explicit dimension reduction and its applications. In *CCC*, pages 262–272, 2011.
12. M. Ledoux and M. Talagrand. *Probability in Banach spaces: isoperimetry and processes*. Springer, 1991.
13. J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.
14. D. Sivakumar. Algorithmic derandomization via complexity theory. In *STOC*, pages 619–626, 2002.
15. D. Zuckerman. Randomness-optimal oblivious sampling. *Random Struct. Algorithms*, 11(4):345–367, 1997.