The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction

Kasper Green Larsen* Jelani Nelson†

November 1, 2014

Abstract

For any $n > 1$ and $0 < \varepsilon < 1/2$, we show the existence of an $n^{O(1)}$-point subset X of $\mathbb{R}^n$ such that any linear map from (X, ℓ_2) to ℓ_2^m with distortion at most $1 + \varepsilon$ must have $m = \Omega(\min\{n, \varepsilon^{-2} \log n\})$. Our lower bound matches the upper bounds provided by the identity matrix and the Johnson-Lindenstrauss lemma [JL84], improving the previous lower bound of Alon [Alo03] by a $\log(1/\varepsilon)$ factor.

1 Introduction

The Johnson-Lindenstrauss lemma [JL84] states the following.

Theorem 1 (JL lemma [JL84, Lemma 1]). *For any N-point subset X of Euclidean space and any $0 < \varepsilon < 1/2$, there exists a map $f : X \to \ell_2^m$ with $m = O(\varepsilon^{-2} \log N)$ such that*

$$\forall x, y \in X,\ (1 - \varepsilon)\|x - y\|_2^2 \le \|f(x) - f(y)\|_2^2 \le (1 + \varepsilon)\|x - y\|_2^2. \quad (1)$$

We henceforth refer to f satisfying (1) as *having the ε-JL guarantee for X* (often we drop mention of ε when understood from context). The JL lemma has found applications in computer science, signal processing (e.g. compressed sensing), statistics, and mathematics. The main idea in algorithmic applications is that one can transform a high-dimensional problem into a low-dimensional one such that an optimal solution to the lower dimensional problem can be lifted to a nearly optimal solution to the original problem. Due to the decreased dimension, the lower dimensional problem requires fewer resources (time, memory, etc.) to solve. We refer the reader to [Ind01, Vem04, Mat08] for a list of further applications.

All known proofs of the JL lemma with target dimension as stated above in fact provide such a map f which is *linear*. This linearity property is important in several applications. For example in the turnstile model of streaming [Mut05], a vector $x \in \mathbb{R}^n$ receives a stream of coordinate-wise updates each of the form $x_i \leftarrow x_i + \Delta$, where $\Delta \in \mathbb{R}$. The goal is to process x using some $m \ll n$ memory. Thus if one wants to perform dimensionality reduction in a stream, which occurs for example in streaming linear algebra applications [CW09], this can be achieved with linear f since $f(x + \Delta \cdot e_i) = f(x) + \Delta \cdot f(e_i)$. In compressed sensing, another application where linearity of f is inherent, one wishes to (approximately) recover (approximately) sparse signals using few linear measurements [Don06, CT05]. The map f sending a signal to the vector containing some fixed set of linear measurements of it is known to allow for good signal recovery as long as f satisfies the JL guarantee for the set of all k-sparse vectors [CT05].

Given the widespread use of dimensionality reduction across several domains, it is a natural and often-asked question whether the JL lemma is tight: does there exist some X of size N such that any such

*Aarhus University. `larsen@cs.au.dk`. Supported by Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation, grant DNRF84.

†Harvard University. `minilek@seas.harvard.edu`. Supported by NSF CAREER award CCF-1350670, NSF grant IIS-1447471, ONR grant N00014-14-1-0632, and a Google Faculty Research Award.

map f must have $m = \Omega(\varepsilon^{-2} \log N)$? The paper [JL84] introducing the JL lemma provided the first lower bound of $m = \Omega(\log N)$ when ε is smaller than some constant. This was improved by Alon [Alo03], who showed that if $X = \{0, e_1, \ldots, e_n\} \subset \mathbb{R}^n$ is the simplex (thus $N = n + 1$) and $0 < \varepsilon < 1/2$, then any JL map f must embed into dimension $m = \Omega(\min\{n, \varepsilon^{-2} \log n / \log(1/\varepsilon)\})$. Note the first term in the min is achieved by the identity map. Furthermore, the $\log(1/\varepsilon)$ term cannot be removed for this particular X since one can use Reed-Solomon codes to obtain embeddings with $m = O(1/\varepsilon^2)$ (superior to the JL lemma) once $\varepsilon \le n^{-\Omega(1)}$ [Alo03] (see [NNW14] for details). Specifically, for this X it is possible to achieve $m = O(\varepsilon^{-2} \min\{\log N, ((\log N)/\log(1/\varepsilon))^2\})$. Note also for this choice of X we can assume that any f is in fact linear. This is because first we can assume $f(0) = 0$ by translation. Then we can form a matrix $A \in \mathbb{R}^{m \times n}$ such that the ith column of A is $f(e_i)$. Then trivially $Ae_i = f(e_i)$ and $A0 = 0 = f(0)$.

The fact that the JL lemma is not optimal for the simplex for small ε begs the question: is the JL lemma suboptimal for all point sets? This is a major open question in the area of dimensionality reduction, and it has been open since the paper of Johnson and Lindenstrauss 30 years ago.

Our Main Contribution: For any $n > 1$ and $0 < \varepsilon < 1/2$, there is an $n^{O(1)}$-point subset X of $\mathbb{R}^n$ such that any embedding f providing the JL guarantee, and where f *is linear*, must have $m = \Omega(\min\{n, \varepsilon^{-2} \log n\})$. In other words, the JL lemma is optimal in the case where f must be linear.

Our lower bound is optimal: the identity map achieves the first term in the min, and the JL lemma provides the second. Our lower bound is only against linear embeddings, but as stated before: (1) all known proofs of the JL lemma give linear f, and (2) for several applications it is important that f be linear.

Before our work there were two possibilities for dimensionality reduction in ℓ_2: (i) the target dimension m could be reduced for all point sets X, at least for small ε as with the simplex using Reed-Solomon codes, or (ii) there is a higher lower bound for some other point set X which is harder than the simplex. Evidence existed to support both possibilities. On the one hand the simplex was the hardest case in many respects: it gave the highest lower bound known on m [Alo03], and it also was a hardest case for the data-dependent upper bound on m of Gordon [Gor88] (involving the gaussian mean width of the normalized difference vectors $X - X$; we will not delve deeper into this topic here). Meanwhile for (ii), random linear maps were the only JL construction we knew for arbitrary X, and such an approach with random maps is known to require $m = \Omega(\min\{n, \varepsilon^{-2} \log N\})$ [JW13, KMN11] (see Remark 1 below for details).

Thus given the previous state of our knowledge, it was not clear which was more likely between worlds (i) and (ii). Our lower bound gives more support to (ii), since we not only rule out further improvements to JL using random linear maps, but rule out improvements using *any* linear map. Furthermore all known methods for efficient dimensionality reduction in ℓ_2 are via linear maps, and thus circumventing our lower bound would require a fundamentally new approach to the problem. We also discuss in Section 4 what would be needed to push our lower bound to apply to non-linear maps.

Remark 1. It is worth noting that the JL lemma is different from the *distributional* JL (DJL) lemma that often appears in the literature, sometimes with the same name (though the lemmas are different!). In the DJL problem, one is given an integer $n > 1$ and $0 < \varepsilon, \delta < 1/2$, and the goal is to provide a distribution $\mathcal{F}$ over maps $f : \ell_2^n \to \ell_2^m$ with m as small as possible such that for any fixed $x \in \mathbb{R}^n$

$$\mathbb{P}_{f \leftarrow \mathcal{F}}(\|f(x)\|_2 \notin [(1-\varepsilon)\|x\|_2, (1+\varepsilon)\|x\|_2]) < \delta.$$

The existence of such $\mathcal{F}$ with small m implies the JL lemma by taking $\delta < 1/\binom{N}{2}$. Then for any $z \in X - X$, a random $f \leftarrow \mathcal{F}$ fails to preserve the norm of z with probability δ. Thus the probability that there exists $z \in X - X$ which f fails to preserve the norm of is at most $\delta \cdot \binom{N}{2} < 1$, by a union bound. In other words, a random map provides the desired JL guarantee with high probability (and in fact this map is chosen completely obliviously of the input vectors).

The optimal m for the DJL lemma when using linear maps is understood. The original paper [JL84] provided a linear solution to the DJL problem with $m = O(\min\{n, \varepsilon^{-2} \log(1/\delta)\})$, and this was later shown to be optimal for the full range of $\varepsilon, \delta \in (0, 1/2)$ [JW13, KMN11]. Thus when δ is set as above, one obtains the $m = O(\varepsilon^{-2} \log N)$ guarantee of the JL lemma. However, this does not imply that the JL lemma is tight. Indeed, it is sometimes possible to obtain smaller m by avoiding the DJL lemma, such as the Reed-Solomon based embedding construction for the simplex mentioned above (which involves zero randomness).

It is also worth remarking that DJL is desirable for one-pass streaming algorithms, since no properties of X are known when the map f is chosen at the beginning of the stream, and thus the DJL lower bounds of [JW13, KMN11] are relevant in this scenario. However when allowed two passes or more, one could imagine estimating various properties of X in the first pass(es) then choosing some linear f more efficiently based on these properties to perform dimensionality reduction in the last pass. The lower bound of our main theorem shows that the target dimension could not be reduced by such an approach.

1.1 Proof overview

For any $n > 1$ and $\varepsilon \in (0, 1/2)$, we prove the existence of $X \subset \mathbb{R}^n$, $|X| = N = O(n^3)$, s.t. if for $A \in \mathbb{R}^{m \times n}$

$$(1 - \varepsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + \varepsilon)\|x\|_2^2 \text{ for all } x \in X, \tag{2}$$

then $m = \Omega(\varepsilon^{-2} \log n) = \Omega(\varepsilon^{-2} \log N)$. Providing the JL guarantee on $X \cup \{0\}$ implies satisfying (2), and therefore also requires $m = \Omega(\varepsilon^{-2} \log N)$. We show such X exists via the probabilistic method, by letting X be the union of all n standard basis vectors together with several independent gaussian vectors. Gaussian vectors were also the hard case in the DJL lower bound proof of [KMN11], though the details were different.

We now give the idea of the lower bound proof to achieve (2). First, we include in X the vectors $e_1, \ldots, e_n$. Then if $A \in \mathbb{R}^{m \times n}$ for $m \le n$ satisfies (2), this forces every column of A to have roughly unit norm. Then by standard results in covering and packing (see Eqn. (5.7) of [Pis89]), there exists some family of matrices $\mathcal{F} \subset \cup_{t=1}^n \mathbb{R}^{t \times n}$, $|\mathcal{F}| = e^{O(n^2 \log n)}$, such that

$$\inf_{\hat{A} \in \mathcal{F} \cap \mathbb{R}^{m \times n}} \|A - \hat{A}\|_F \le \frac{1}{n^C} \tag{3}$$

for $C > 0$ a constant as large as we like, where $\|\cdot\|_F$ denotes Frobenius norm. Also, by a theorem of Latała [Lat99], for any $\hat{A} \in \mathcal{F}$ and for a random gaussian vector g,

$$\mathbb{P}_g(|\|\hat{A}g\|_2^2 - \operatorname{tr}(\hat{A}^T \hat{A})| \ge \Omega(\sqrt{\log(1/\delta)} \cdot \|\hat{A}^T \hat{A}\|_F)) \ge \delta \tag{4}$$

for any $0 < \delta < 1/2$, where $\operatorname{tr}(\cdot)$ is trace. This is a (weaker version of the) statement that for gaussians, the Hanson-Wright inequality [HW71] not only provides an upper bound on the tail of degree-two gaussian chaos, but also is a lower bound. (The strong form of the previous sentence, without the parenthetical qualifier, was proven in [Lat99], but we do not need this stronger form for our proof – essentially the difference is that in stronger form, (4) is replaced with a stronger inequality also involving the operator norm $\|\hat{A}^T \hat{A}\|$.)

It also follows by standard gaussian concentration that a random gaussian vector g satisfies

$$\mathbb{P}_g(|\|g\|_2^2 - n| > C\sqrt{n \log(1/\delta)}) < \delta/2 \tag{5}$$

Thus by a union bound, the events of (4), (5) happen simultaneously with probability $\Omega(\delta)$. Thus if we take N random gaussian vectors, the probability that the events of (4), (5) never happen simultaneously for any of the N gaussians is at most $(1 - \Omega(\delta))^N = e^{-\Omega(\delta N)}$. By picking N sufficiently large and $\delta = 1/\text{poly}(n)$, a union bound over $\mathcal{F}$ shows that for every $\hat{A} \in \mathcal{F}$, one of the N gaussians satisfies the events of (4) and (5) simultaneously. Specifically, there exist $N = O(n^3)$ vectors $\{v_1, \ldots, v_N\} = V \subset \mathbb{R}^n$ such that

- Every $v \in V$ has $\|v\|_2^2 = n \pm O(\sqrt{n \lg n})$
- For any $\hat{A} \in \mathcal{F}$ there exists some $v \in V$ such that $|\|\hat{A}v\|_2^2 - \operatorname{tr}(\hat{A}^T \hat{A})| = \Omega(\sqrt{\lg n} \cdot \|\hat{A}\|_F)$.

The final definition of X is $\{e_1, \ldots, e_n\} \cup V$. Then, using (2) and (3), we show that the second bullet implies

$$\operatorname{tr}(\hat{A}^T \hat{A}) = n \pm O(\varepsilon n), \text{ and } |\|Av\|_2^2 - n| = \Omega(\sqrt{\ln n} \cdot \|\hat{A}^T \hat{A}\|_F) - O(\varepsilon n). \tag{6}$$

where $\pm B$ represents a value in $[-B, B]$. But then by the triangle inequality, the first bullet above, and (2),

$$\left|\|Av\|_2^2 - n\right| \le \left|\|Av\|_2^2 - \|v\|_2^2\right| + \left|\|v\|_2^2 - n\right| = O(\varepsilon n + \sqrt{n \lg n}). \tag{7}$$

Combining (6) and (7) implies

$$\operatorname{tr}(\hat{A}^T \hat{A}) = \sum_{i=1}^{n} \hat{\lambda}_i \ge (1 - O(\varepsilon))n, \text{ and } \|\hat{A}^T \hat{A}\|_F^2 = \sum_{i=1}^{n} \hat{\lambda}_i^2 = O\left(\frac{\varepsilon^2 n^2}{\log n} + n\right)$$

where $(\hat{\lambda}_i)$ are the eigenvalues of $\hat{A}^T \hat{A}$. With bounds on $\sum_i \hat{\lambda}_i$ and $\sum_i \hat{\lambda}_i^{\ 2}$ in hand, a lower bound on $\operatorname{rank}(\hat{A}^T \hat{A}) \le m$ follows by Cauchy-Schwarz (this last step is also common to the proof of [Alo03]).

Remark 2. It is not crucial in our proof that N be proportional to n^3. Our techniques straightforwardly extend to show that N can be any value which is $\Omega(n^{2+\gamma})$ for any constant $\gamma > 0$.

2 Preliminaries

Henceforth a *standard gaussian* random variable $g \in \mathbb{R}$ is a gaussian with mean 0 and variance 1. If we say $g \in \mathbb{R}^n$ is standard gaussian, then we mean that g is a multivariate gaussian with identity covariance matrix (i.e. its entries are independent standard gaussian). Also, the notation $\pm B$ denotes a value in $[-B, B]$. For a real matrix $A = (a_{i,j})$, $\|A\|$ is the $\ell_2 \to \ell_2$ operator norm, and $\|A\|_F = (\sum_{i,j} a_{i,j}^2)^{1/2}$ is Frobenius norm.

In our proof we depend on some previous work. The first theorem is due to Latała [Lat99] and says that, for gaussians, the Hanson-Wright inequality is not only an upper bound but also a lower bound.

Theorem 2 ([Lat99, Corollary 2])**.** *There exists universal $c > 0$ such that for $g \in \mathbb{R}^n$ standard gaussian and $A = (a_{i,j})$ an $n \times n$ real symmetric matrix with zero diagonal,*

$$\forall t \ge 1, \ \mathop{\mathbb{P}}_{g}\left(|g^T A g| > c(\sqrt{t} \cdot \|A\|_F + t \cdot \|A\|)\right) \ge \min\{c, e^{-t}\}$$

Theorem 2 implies the following corollary.

Corollary 1. *Let g, A be as in Theorem 2, but where A is no longer restricted to have zero diagonal. Then*

$$\forall t \ge 1, \ \mathop{\mathbb{P}}_{g}\left(|g^T A g - tr(A)| > c(\sqrt{t} \cdot \|A\|_F + t \cdot \|A\|)\right) \ge \min\{c, e^{-t}\}$$

Proof. Let N be a positive integer. Define $\tilde{g} = (\tilde{g}_{1,1}, \tilde{g}_{1,2}, \ldots, \tilde{g}_{1,N}, \ldots, \tilde{g}_{n,1}, \tilde{g}_{n,2}, \ldots, \tilde{g}_{n,N})$ a standard gaussian vector. Then g_i is equal in distribution to $N^{-1/2} \sum_{j=1}^{N} \tilde{g}_{i,j}$. Define $\tilde{A}_N$ as the $nN \times nN$ matrix formed by converting each entry $a_{i,j}$ of A into an $N \times N$ block with each entry being $a_{i,j}/N$. Then

$$g^T A g - \operatorname{tr}(A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} g_i g_j - \operatorname{tr}(A) \stackrel{d}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{N} \sum_{s=1}^{N} \frac{a_{i,j}}{N} \tilde{g}_{i,r} \tilde{g}_{j,s} - \operatorname{tr}(A) \stackrel{\text{def}}{=} \tilde{g}^T \tilde{A}_N \tilde{g} - \operatorname{tr}(\tilde{A}_N)$$

where $\stackrel{d}{=}$ denotes equality in distribution (note $\operatorname{tr}(A) = \operatorname{tr}(\tilde{A}_N)$). By the weak law of large numbers

$$\forall \lambda > 0, \ \lim_{N \to \infty} \mathbb{P}\left(|\tilde{g}^T \tilde{A}_N \tilde{g} - \operatorname{tr}(\tilde{A}_N)| > \lambda\right) = \lim_{N \to \infty} \mathbb{P}\left(|\tilde{g}^T (\tilde{A}_N - \tilde{D}_N) \tilde{g}| > \lambda\right) \tag{8}$$

where $\tilde{D}_N$ is diagonal containing the diagonal elements of $\tilde{A}_N$. Note $\|\tilde{A}_N\| = \|A\|$. This follows since if we have the singular value decomposition $A = \sum_i \sigma_i u_i v_i^T$ (where the $\{u_i\}$ and $\{v_i\}$ are each orthonormal,

$\sigma_i > 0$, and $\|A\|$ is the largest of the σ_i), then $\tilde{A}_N = \sum_i \sigma_i u_i^{(N)} (v_i^{(N)})^T$ where $u_i^{(N)}$ is equal to u_i but where every coordinate is replicated N times and divided by $\sqrt{N}$. This implies $|\|\tilde{A}_N - \tilde{D}_N\| - \|A\|| \le \|\tilde{D}_N\| = \max_i |a_{i,i}|/N = o_N(1)$ by the triangle inequality. Therefore $\lim_{N\to\infty} \|\tilde{A}_N - \tilde{D}_N\| = \|A\|$. Also $\lim_{N\to\infty} \|\tilde{A}_N - \tilde{D}_N\|_F = \|A\|_F$. Our corollary follows by applying Theorem 2 to the right side of (8). □

The next lemma follows from gaussian concentration of Lipschitz functions [Pis86, Corollary 2.3]. It also follows directly from the Hanson-Wright inequality [HW71].

Lemma 1. *For some universal $c > 0$ and $g \in \mathbb{R}^n$ a standard gaussian, $\forall t > 0$ $\mathbb{P}(|\|g\|_2^2 - n| > c\sqrt{nt}) < e^{-t}$.*

The following corollary summarizes the above in a form that will be useful later.

Corollary 2. *For $A \in \mathbb{R}^{d\times n}$ let $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ be the eigenvalues of A^TA. Let $g^{(1)}, \ldots, g^{(N)} \in \mathbb{R}^n$ be independent standard gaussian vectors. For some universal constants $c_1, c_2, \delta_0 > 0$ and any $0 < \delta < \delta_0$*

$$\mathbb{P}\left(\nexists j \in [N] : \left\{ \left| \|Ag^{(j)}\|_2^2 - \sum_{i=1}^n \lambda_i \right| \ge c_1\sqrt{\ln(1/\delta)} \left(\sum_{i=1}^n \lambda_i^2\right)^{1/2} \right\} \wedge \left\{ |\|g^{(j)}\|_2^2 - n| \le c_2\sqrt{n\ln(1/\delta)} \right\}\right) \le e^{-N\delta}. \tag{9}$$

Proof. We will show that for any fixed $j \in [N]$ it holds that

$$\mathbb{P}\left(\left\{ \left| \|Ag^{(j)}\|_2^2 - \sum_{i=1}^n \lambda_i \right| \ge c_1\sqrt{\ln(1/\delta)} \left(\sum_{i=1}^n \lambda_i^2\right)^{1/2} \right\} \wedge \left\{ \|g^{(j)}\|_2^2 \le n + c_2\sqrt{n\ln(1/\delta)} \right\}\right) > \delta \tag{10}$$

Then, since the g_j are independent, the left side of (9) is at most $(1-\delta)^N \le e^{-\delta N}$.

Now we must show (10). It suffices to show that

$$\mathbb{P}\left(|\|g^{(j)}\|_2^2 - n| \le c_2\sqrt{n\ln(1/\delta)}\right) > 1 - \delta/2 \tag{11}$$

and

$$\mathbb{P}\left(\left| \|Ag^{(j)}\|_2^2 - \sum_{i=1}^n \lambda_i \right| \ge c_1\sqrt{\ln(1/\delta)} \left(\sum_{i=1}^n \lambda_i^2\right)^{1/2}\right) > \delta/2 \tag{12}$$

since (10) would then follow from a union bound. Eqn. (11) follows immediately from Lemma 1 for c_2 chosen sufficiently large. For Eqn. (12), note $\|Ag^{(j)}\|_2^2 = g^TA^TAg$. Then $\sum_i \lambda_i = \mathrm{tr}(A^TA)$ and $(\sum_i \lambda_i^2)^{1/2} = \|A^TA\|_F$. Then (12) frollows from Corollary 1 for δ smaller than some sufficiently small constant δ_0. □

We also need a standard estimate on entropy numbers (covering the unit ℓ_∞^{mn} ball by ℓ_2^{mn} balls).

Lemma 2. *For any parameter $0 < \alpha < 1$, there exists a family $\mathcal{F}_\alpha \subseteq \bigcup_{m=1}^n \mathbb{R}^{m\times n}$ of matrices with the following two properties:*

1. *For any matrix $A \in \bigcup_{m=1}^n \mathbb{R}^{m\times n}$ having all entries bounded in absolute value by 2, there is a matrix $\hat{A} \in \mathcal{F}_\alpha$ such that A and $\hat{A}$ have the same number of rows and $B = A - \hat{A}$ satisfies $tr(B^TB) \le \alpha/100$.*

2. *$|\mathcal{F}_\alpha| = e^{O(n^2\ln(n/\alpha))}$.*

Proof. We construct $\mathcal{F}_\alpha$ as follows: For each integer $1 \le m \le n$, add all $m \times n$ matrices having entries of the form $i\frac{\sqrt{\alpha}}{10n}$ for integers $i \in \{-20n/\sqrt{\alpha}, \ldots, 20n/\sqrt{\alpha}\}$. Then for any matrix $A \in \bigcup_{m=1}^n \mathbb{R}^{m\times n}$ there is a matrix $\hat{A} \in \mathcal{F}_\alpha$ such that A and $\hat{A}$ have the same number of rows and every entry of $B = A - \hat{A}$ is bounded in absolute value by $\frac{\sqrt{\alpha}}{10n}$. This means that every diagonal entry of B^TB is bounded by $n\alpha/(100n^2)$ and thus $\mathrm{tr}(B^TB) \le \alpha/100$. The size of $\mathcal{F}_\alpha$ is bounded by $n(40n/\sqrt{\alpha})^{n^2} = e^{O(n^2\ln(n/\alpha))}$. □

3 Proof of main theorem

Lemma 3. *Let $\mathcal{F}_\alpha$ be as in Lemma 2 with $1/\operatorname{poly}(n) \le \alpha < 1$. Then there exists a set of $N = O(n^3)$ vectors $v_1, \dots, v_N$ in $\mathbb{R}^n$ such that for every matrix $A \in \mathcal{F}_\alpha$, there is an index $j \in [N]$ such that*

(i) $|\|Av_j\|_2^2 - \sum_i \lambda_i| = \Omega\left(\sqrt{\ln n \sum_i \lambda_i^2}\right)$.

(ii) $|\|v_j\|_2^2 - n| = O(\sqrt{n \ln n})$.

Proof. Let $g^{(1)}, \dots, g^{(N)} \in \mathbb{R}^n$ be independent standard gaussian. Let $A \in \mathcal{F}_\alpha$ and apply Corollary 2 with $\delta = n^{-1/4} = N^{-1/12}$. With probability $1 - e^{-\Omega(n^{3-1/4})}$, one of the $g^{(j)}$ for $j \in [N]$ satisfies (i) and (ii) for A. Since $|\mathcal{F}_\alpha| = e^{O(n^2 \ln(n/\alpha))}$, the claim follows by a union bound over all matrices in $\mathcal{F}_\alpha$. $\square$

Theorem 3. *For any $0 < \varepsilon < 1/2$, there exists a set $V \subset \mathbb{R}^n$, $|V| = N = n^3 + n$, such that if A is a matrix in $\mathbb{R}^{m \times n}$ satisfying $\|Av_i\|_2^2 \in (1 \pm \varepsilon)\|v_i\|_2^2$ for all $v_i \in V$, then $m = \Omega(\min\{n, \varepsilon^{-2} \lg n\})$.*

Proof. We can assume $\varepsilon > 1/\sqrt{n}$ since otherwise an $m = \Omega(n)$ lower bound already follows from [Alo03]. To construct V, we first invoke Lemma 3 with $\alpha = \varepsilon^2/n^2$ to find n^3 vectors $w_1, \dots, w_{n^3}$ such that for all matrices $\tilde{A} \in \mathcal{F}_{\varepsilon^2/n^2}$, there exists an index $j \in [n^3]$ for which:

1. $|\|\tilde{A}w_j\|_2^2 - \sum_i \tilde{\lambda}_i| \ge \Omega\left(\sqrt{(\ln n)\sum_i \tilde{\lambda}_i^2}\right)$.
2. $|\|w_j\|_2^2 - n| = O(\sqrt{n \ln n})$.

where $\tilde{\lambda}_1 \ge \dots \ge \tilde{\lambda}_n \ge 0$ denote the eigenvalues of $\tilde{A}^T\tilde{A}$. We let $V = \{e_1, \dots, e_n, w_1, \dots, w_{n^3}\}$ and claim this set of $N = n^3 + n$ vectors satisfies the theorem. Here e_i denotes the i'th standard unit vector.

To prove this, let $A \in \mathbb{R}^{m \times n}$ be a matrix with $m \le n$ satisfying $\|Av\|_2^2 \in (1 \pm \varepsilon)\|v\|_2^2$ for all $v \in V$. Now observe that since $e_1, \dots, e_n \in V$, A satisfies $\|Ae_i\|_2^2 \in (1 \pm \varepsilon)\|e_i\|_2^2 = (1 \pm \varepsilon)$ for all e_i. Hence all entries $a_{i,j}$ of A must have $a_{i,j}^2 \le (1 + \varepsilon) < 2$ (and in fact, all columns of A have ℓ_2 norm at most $\sqrt{2}$). This implies that there is an $m \times n$ matrix $\hat{A} \in \mathcal{F}_{\varepsilon^2/n^2}$ such that $B = A - \hat{A} = (b_{i,j})$ satisfies $\operatorname{tr}(B^TB) \le \varepsilon^2/(100n^2)$. Since $\operatorname{tr}(B^TB) = \|B\|_F^2$, this also implies $\|B\|_F \le \varepsilon/(10n)$. Then by Cauchy-Schwarz,

$$\begin{aligned}
\sum_{i=1}^n \hat{\lambda}_i &= \operatorname{tr}(\hat{A}^T\hat{A}) \\
&= \operatorname{tr}((A-B)^T(A-B)) \\
&= \operatorname{tr}(A^TA) + \operatorname{tr}(B^TB) - \operatorname{tr}(A^TB) - \operatorname{tr}(B^TA) \\
&= \sum_{i=1}^n \|Ae_i\|_2^2 + \operatorname{tr}(B^TB) - \operatorname{tr}(A^TB) - \operatorname{tr}(B^TA) \\
&= n \pm (O(\varepsilon n) + 2n \cdot \max_j(\sum_i b_{i,j}^2)^{1/2} \cdot \max_k(\sum_i a_{i,k}^2)^{1/2}) \\
&= n \pm (O(\varepsilon n) + 2n \cdot \|B\|_F \cdot \sqrt{2}) \\
&= n \pm O(\varepsilon n).
\end{aligned}$$

Thus from our choice of V there exists a vector $v^* \in V$ such that

(i) $|\|\hat{A}v^*\|_2^2 - n| \ge \Omega\left(\sqrt{(\ln n)\sum_i \hat{\lambda}_i^2}\right) - O(\varepsilon n)$.

(ii) $|\|v^*\|_2^2 - n| = O(\sqrt{n \ln n})$.

Note $\|B\|^2 \le \|B\|_F^2 = \operatorname{tr}(B^TB) \le \varepsilon^2/(100n^2)$ and $\|\hat{A}\|^2 \le \|\hat{A}\|_F^2 \le (\|A\|_F + \|B\|_F)^2 = O(n^2)$. Then by (i)

(iii)

$$\begin{aligned}
|\|Av^*\|_2^2 - n| &= |\|\hat{A}v^*\|_2^2 + \|Bv^*\|_2^2 + 2\langle \hat{A}v^*, Bv^*\rangle - n| \\
&\geq \Omega\left(\sqrt{(\ln n)\sum_i \hat{\lambda}_i^2}\right) - \|Bv^*\|_2^2 - 2|\langle \hat{A}v^*, Bv^*\rangle| - O(\varepsilon n) \\
&\geq \Omega\left(\sqrt{(\ln n)\sum_i \hat{\lambda}_i^2}\right) - \|B\|^2 \cdot \|v^*\|_2^2 - 2\|B\| \cdot \|A\| \cdot \|v^*\|_2^2 - O(\varepsilon n) \\
&= \Omega\left(\sqrt{(\ln n)\sum_i \hat{\lambda}_i^2}\right) - O(\varepsilon n).
\end{aligned}$$

We assumed $|\|Av^*\|_2^2 - \|v^*\|_2^2| = O(\varepsilon\|v^*\|_2^2) = O(\varepsilon n)$. Therefore by (ii),

$$\begin{aligned}
\left|\|Av^*\|_2^2 - n\right| &\leq \left|\|Av^*\|_2^2 - \|v^*\|_2^2\right| + \left|\|v^*\|_2^2 - n\right| \\
&= O(\varepsilon n + \sqrt{n \ln n}),
\end{aligned}$$

which when combined with (iii) implies

$$\sum_{i=1}^n \hat{\lambda}_i^2 = O\left(\frac{\varepsilon^2 n^2}{\ln n} + n\right).$$

To complete the proof, by Cauchy-Schwarz since exactly $\mathrm{rank}(\hat{A}^T\hat{A})$ of the $\hat{\lambda}_i$ are non-zero,

$$\frac{n^2}{2} \leq \left(\sum_{i=1}^n \hat{\lambda}_i\right)^2 \leq \mathrm{rank}(\hat{A}^T\hat{A}) \cdot \left(\sum_{i=1}^n \hat{\lambda}_i^{\,2}\right) \leq m \cdot O\left(\frac{\varepsilon^2 n^2}{\ln n} + n\right)$$

Rearranging gives $m = \Omega(\min\{n, \varepsilon^{-2}\ln n\}) = \Omega(\min\{n, \varepsilon^{-2}\ln N\})$ as desired.

□

4 Discussion

One obvious future direction is to obtain an $m = \Omega(\min\{n, \varepsilon^{-2}\log N\})$ lower bound that also applies to non-linear maps. Our hard set X contains $N = O(n^3)$ points in $\mathbb{R}^n$ (though as remarked earlier, our techniques easily imply $N = O(n^{2+\gamma})$ points suffice). Any embedding f could be assumed linear without loss of generality if the elements of X were linearly independent, but clearly this cannot happen if $N > n$ (as is the case for our X). Thus a first step toward a lower bound against non-linear embeddings is to obtain a hard X with $N = |X|$ as small as possible. One step in this direction could be the following. Observe that our lower bound only uses that $\|f(x)\|_2 \approx \|x\|_2$ for each $x \in X$, whereas the full JL lemma requires that all distance vectors $X - X$ have their norms preserved. Thus one could hope to exploit this fact and take $|X| = \Theta(n^{1+\gamma})$, say, since then $X - X$ would still have the $\Theta(n^{2+\gamma})$ points needed to carry out the union bound of Lemma 3. The problem is that these $\Theta(n^{2+\gamma})$ points would not be independent, and thus the argument of Corollary 2 would not apply. A more careful argument would have to be crafted. Of course, one would still need a further idea to then reduce N from $\Theta(n^{1+\gamma})$ down to n.

Acknowledgments

We thank Radosław Adamczak for pointing out how to derive Corollary 1 from Theorem 2, and for pointing out the reference [AW13], which uses a more involved but similar argument.

References

[Alo03] Noga Alon. Problems and results in extremal combinatorics–I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[AW13] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *CoRR*, abs/1304.1826, 2013.

[CT05] Emmanuel Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

[CW09] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, (STOC)*, pages 205–214, 2009.

[Don06] David Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[Gor88] Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 1971.

[Ind01] Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 10–33, 2001.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[JW13] T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26, 2013.

[KMN11] Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Proceedings of the 15th International Workshop on Randomization and Computation (RANDOM)*, pages 628–639, 2011.

[Lat99] Rafał Latała. Tail and moment estimates for some types of chaos. *Studia Math.*, 135:39–53, 1999.

[Mat08] Jirí Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[Mut05] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[NNW14] Jelani Nelson, Huy L. Nguyễn, and David P. Woodruff. On deterministic sketching and streaming for sparse recovery and norm estimation. *Linear Algebra and its Applications, Special Issue on Sparse Approximate Solution of Linear Systems*, 441:152–167, 2014.

[Pis86] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. *Probability and Analysis, Lecture Notes in Math.*, 1206:167–241, 1986.

[Pis89] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1989.

[Vem04] Santosh Vempala. *The random projection method*, volume 65 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, 2004.