Optimality of the Johnson-Lindenstrauss lemma

Kasper Green Larsen* Jelani Nelson†

November 8, 2017

Abstract

For any $d, n \geq 2$ and $1/(\min\{n, d\})^{0.4999} < \varepsilon < 1$, we show the existence of a set of n vectors $X \subset \mathbb{R}^d$ such that any embedding $f : X \to \mathbb{R}^m$ satisfying

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

must have

$$m = \Omega(\varepsilon^{-2} \lg n).$$

This lower bound matches the upper bound given by the Johnson-Lindenstrauss lemma [JL84]. Furthermore, our lower bound holds for nearly the full range of ε of interest, since there is always an isometric embedding into dimension $\min\{d, n\}$ (either the identity map, or projection onto *span*(X)).

Previously such a lower bound was only known to hold against *linear* maps f, and not for such a wide range of parameters ε, n, d [LN16]. The best previously known lower bound for general f was $m = \Omega(\varepsilon^{-2} \lg n / \lg(1/\varepsilon))$ [Wel74, Lev83, Alo03], which is suboptimal for any $\varepsilon = o(1)$.

1 Introduction

In modern algorithm design, often data is high-dimensional, and one seeks to first pre-process the data via some *dimensionality reduction* scheme that preserves geometry in such a way that is acceptable for particular applications. The lower-dimensional embedded data has the benefit of requiring less storage, less communication bandwith to be transmitted over a network, and less time to be analyzed by later algorithms. Such schemes have been applied to good effect in a diverse range of areas, such as streaming algorithms [Mut05], numerical linear algebra [Woo14], compressed sensing [CRT06, Don06], graph sparsification [SS11], clustering [BZMD15, CEM+15], nearest neighbor search [HIM12], and many others.

A cornerstone dimensionality reduction result is the following *Johnson-Lindenstrauss (JL) lemma* [JL84].

Theorem 1 (JL lemma)**.** *Let $X \subset \mathbb{R}^d$ be any set of size n, and let $\varepsilon \in (0, 1/2)$ be arbitrary. Then there exists a map $f : X \to \mathbb{R}^m$ for some $m = O(\varepsilon^{-2} \lg n)$ such that*

$$\forall x, y \in X,\ (1-\varepsilon)\|x-y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2. \tag{1}$$

Even though the JL lemma has found applications in a plethora of different fields over the past three decades, its optimality has still not been settled. In the original paper by Johnson and Lindenstrauss [JL84], it was proven that for any $\varepsilon < 1/2$, there exists n point sets $X \subset \mathbb{R}^n$ for which any embedding $f : X \to \mathbb{R}^m$ providing (1) must have $m = \Omega(\lg n)$. This was later improved in [Lev83, Alo03], which showed the existence of an n point set $X \subset \mathbb{R}^n$, such that any f providing (1) must have $m = \Omega(\min\{n, \varepsilon^{-2} \lg n / \lg(1/\varepsilon)\})$, which falls short of the JL lemma for any $\varepsilon = o(1)$. This lower bound can also be obtained from the Welch bound [Wel74], which states $\varepsilon^{2k} \geq (1/(n-1))(n/\binom{m+k-1}{k} - 1)$ for any positive integer k, by choosing $2k = \lceil \lg n / \lg(1/\varepsilon) \rceil$. The lower bound can also be extended to hold for any $n \leq e^{c\varepsilon^2 d}$ for some constant $c > 0$.

*Aarhus University. larsen@cs.au.dk. Supported by Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation, grant DNRF84, a Villum Young Investigator Grant and an AUFF Starting Grant.

†Harvard University. minilek@seas.harvard.edu. Supported by NSF CAREER award CCF-1350670, NSF grant IIS-1447471, ONR Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

Our Contribution In this paper, we finally settle the optimality of the JL lemma. Furthermore, we do so for almost the full range of ε.

Theorem 2. *For any integers $n, d \geq 2$ and $\varepsilon \in (\lg^{0.5001} n/\sqrt{\min\{n,d\}}, 1)$, there exists a set of points $X \subset \mathbb{R}^d$ of size n, such that any map $f : X \to \mathbb{R}^m$ providing the guarantee* (1) *must have*

$$m = \Omega(\varepsilon^{-2} \lg(\varepsilon^2 n)). \tag{2}$$

Here it is worth mentioning that the JL lemma can be used to give an upper bound of

$$m = O(\min\{n, d, \varepsilon^{-2} \lg n\}),$$

where the d term is obvious (the identity map) and the n term follows by projecting onto the $\leq n$-dimensional subspace spanned by X. Thus a requirement of at least $\varepsilon = \Omega(1/\sqrt{\min\{n,d\}})$ is certainly necessary for the lower bound (2) to be true, which our constraint on ε matches up to the $\lg^{0.5001} n$ factor.

We also make the following conjecture concerning the behavior of the optimal form of Euclidean dimension reduction possible as $\varepsilon \to 1/\sqrt{\min\{n,d\}}$. Note the $\lg(\varepsilon^2 n)$ term as opposed to $\lg n$ in the upper bound.

Conjecture 1. *If $f(n, d, \varepsilon)$ denotes the smallest m such that all n-point subsets of ℓ_2^d can be embedded into ℓ_2^m with distortion at most $1+\varepsilon$, then for all $n, d > 1$ and $0 < \varepsilon < 1$, $f(n, d, \varepsilon) = \Theta(\min\{n, d, \varepsilon^{-2} \lg(2+\varepsilon^2 n)\})$.*

It is worth mentioning that the arguments in previous work [Wel74, Alo03, LN16] all produced hard point sets P which were nearly orthogonal so that any embedding into an *incoherent* collection provided low distortion under the Euclidean metric. Recall P is ε-incoherent if every $x \in P$ has unit ℓ_2 norm, and $\forall x \neq y \in P$ one has $|\langle x, y\rangle| = O(\varepsilon)$. Unfortunately though, it is known that for any $\varepsilon < 2^{-\omega(\sqrt{\lg n})}$, an incoherent collection of n vectors in dimension $m = o(\varepsilon^{-2} \lg n)$ exists, beating the guarantee of the JL lemma. The construction is based on Reed-Solomon codes (see for example [AGHP92, NNW14]). Thus proving Theorem 2 requires a very different construction of a hard point set when compared with previous work.

1.1 Prior Work

Prior to our work, a result of the authors [LN16] showed an $m = \Omega(\varepsilon^{-2} \lg n)$ bound in the restricted setting where f must be *linear*. This left open the possibility that the JL lemma could be improved upon by making use of nonlinear embeddings. Indeed, as mentioned above even the hard instance of [LN16] enjoys the existence of a nonlinear embedding into $m = o(\varepsilon^{-2} \lg n)$ dimension for $\varepsilon < 2^{-\omega(\sqrt{\lg n})}$. Furthermore, that result only provided hard instances with $n \leq poly(d)$, and furthermore n had to be sufficiently large (at least $\Omega(d^{1+\gamma}/\varepsilon^2)$ for any constant $\gamma > 0$).

Also related is the so-called *distributional JL* (DJL) lemma. The original proof of the JL lemma in [JL84] is via *random projection*, i.e. ones picks a uniformly random rotation U then defines $f(x)$ to be the projection of Ux onto its first m coordinates, scaled by $1/\sqrt{m}$ in order to have the correct squared Euclidean norm in expectation. Note that this construction of f is both *linear*, and *oblivious* to the data set X. Indeed, all known proofs of the JL lemma proceed by instantiating distributions $\mathcal{D}_{\varepsilon,\delta}$ satisfying the guarantee of the below distributional JL (DJL) lemma.

Lemma 1 (Distributional JL (DJL) lemma)**.** *For any integer $d \geq 1$ and any $0 < \varepsilon, \delta < 1/2$, there exists a distribution $\mathcal{D}_{\varepsilon,\delta}$ over $m \times d$ real matrices for some $m \lesssim \varepsilon^{-2} \lg(1/\delta)$ such that*

$$\forall u \in \mathbb{R}^d, \ \mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}}(|\|\Pi u\|_2 - \|u\|_2| > \varepsilon \|u\|_2) < \delta. \tag{3}$$

One then proves the JL lemma by proving the DJL lemma with $\delta < 1/\binom{n}{2}$, then performing a union bound over all $u \in \{x - y : x, y \in X\}$ to argue that Π simultaneously preserves all norms of such difference vectors simultaneously with positive probability. It is known that the DJL lemma is tight [JW13, KMN11]; namely any distribution $\mathcal{D}_{\varepsilon,\delta}$ over $\mathbb{R}^{m\times n}$ satisfying (3) must have $m = \Omega(\min\{d, \varepsilon^{-2} \lg(1/\delta)\})$. Note though that, prior to our current work, it may have been possible to improve upon the JL lemma by avoiding the DJL lemma. Our main result implies that, unfortunately, this is not the case: obtaining (1) via the DJL lemma combined with a union bound is optimal.

1.2 Subsequent Work

After the initial dissemination of this work, Alon and Klartag asked the question of the optimal space complexity for solving the static "approximate dot product" problem on the sphere in d dimensions [AK17]. In this problem one is given a set P of n points $x_1, \ldots, x_n$ in S^{d-1} to preprocess into a data structure, as well as an error parameter ε. Then in response to $\texttt{query}(i, j)$, one must output $\langle x_i, x_j \rangle$ with additive error at most ε. The work [KOR00] provides a solution using space $O(\varepsilon^{-2} n \lg n)$ bits, which turns out to be optimal iff $d = \Omega(\varepsilon^{-2} \lg n)$, shown by [AK17]. In fact [AK17] was able to provide an understanding of the precise asymptotic space complexity $s(n, d, \varepsilon)$ of this problem for all ranges of n, d, ε. This understanding as a consequence provides an alternate proof of the optimality of the JL lemma, since their work implies $s(n, n, 2\varepsilon) \gg s(n, c\varepsilon^{-2} \lg n, \varepsilon)$ for $c > 0$ a small constant (and if dimension-reduction into dimension d' were always possible, one would have $s(n, n, 2\varepsilon) \leq s(n, d', \varepsilon)$ by first dimension-reducing the input!).

In terms of proof methods, unlike [Alo03, Wel74], our work uses an encoding argument. We proceed in a somewhat ad hoc fashion, showing that one can use simple upper bounds on the sizes of ε-nets of various convex bodies to conclude that dimension reduction far below the JL upper bound would imply an encoding scheme that is too efficient to exist for some task, based on rounding vectors to net points (see Section 3 for an overview). Interestingly enough, the original $m = \Omega(\lg n)$ lower bound of [JL84] was via a volumetric argument, which is related to the packing and covering bounds one needs to execute our encoding argument! The work of [AK17] on understanding $s(n, d, \varepsilon)$ is also via an encoding argument. They observe that the question of understanding $s(n, d, \varepsilon)$ is essentially equivalent to understanding the logarithm of the optimal size of an ε-net under entrywise ℓ_∞ norm of $n \times n$ Gram matrices of rank d, since P can be encoded as the name of the closest point in the net to its Gram matrix. They then proceed to provide tight upper and lower bounds on the optimal net size for the full range of parameters.

The work [AK17] also made progress toward Conjecture 1. In particular, they proved the lower bound for all ranges of parameters, thus removing the "$\lg^{0.5001} n$" term in our requirement on ε in Theorem 2. As for the upper bound, they made progress on a *bipartite* version of the conjecture. In particular, they showed that for any $2n$ vectors $x_1, \ldots, x_n, y_1, \ldots, y_n \in S^{d-1}$, one can find $2n$ vectors $a_1, \ldots, a_n, b_1, \ldots, b_n \in S^{m-1}$ for $m = O(\varepsilon^{-2} \lg(2 + \varepsilon^2 n))$ so that for all $i, j \in [n]$, $|\langle x_i, y_j \rangle - \langle a_i, b_j \rangle| < \varepsilon$. No promise is given for dot product preservation amongst the x_i's internally, or amongst the y_j's internally. Also note that dot product preservation up to additive ε error does not always imply norm preservation with relative error $1 + \varepsilon$, i.e. when distances are small.

2 Preliminaries on Covering Convex Bodies

We here state a standard result on covering numbers. The proof is via a volume comparison argument; see for example [Pis89, Equation (5.7)].

Lemma 2. *Let E be an m-dimensional normed space, and let B_E denote its unit ball. For any $0 < \varepsilon < 1$, one can cover B_E using at most $2^{m \lg(1+2/\varepsilon)}$ translated copies of εB_E.*

Corollary 1. *Let T be an origin symmetric convex body in $\mathbb{R}^m$. For any $0 < \varepsilon < 1$, one can cover T using at most $2^{m \lg(1+2/\varepsilon)}$ translated copies of εT.*

Proof. The Minkowski functional of an origin symmetric convex body T, when restricted to the subspace spanned by vectors in T, is a norm for which T is the unit ball (see e.g. [Tho96, Proposition 1.1.8]). It thus follows from Lemma 2 that T can be covered using at most $2^{m \lg(1+2/\varepsilon)}$ translated copies of εT. □

In the remainder of the paper, we often use the notation B_p^d to denote the unit ℓ_p ball in $\mathbb{R}^d$.

3 Lower Bound Proof

In the following, we start by describing the overall strategy in our proof. This first gives a fairly simple proof of a sub-optimal lower bound. We then introduce the remaining ideas needed and complete the full proof.

The proof goes via a counting argument. More specifically, we construct a large family $\mathcal{P} = \{P_1, P_2, \dots\}$ of very different sets of n points in $\mathbb{R}^d$. We then assume all point sets in $\mathcal{P}$ can be embedded into $\mathbb{R}^m$ while preserving all pairwise distances to within $(1+\varepsilon)$. Letting $f_1(P_1), f_2(P_2), \dots,$ denote the embedded point sets, we then argue that our choice of $\mathcal{P}$ ensures that any two $f_i(P_i)$ and $f_j(P_j)$ must be very different. If m is too low, this is impossible as there are not enough sufficiently different point sets in $\mathbb{R}^m$.

In greater detail, the point sets in $\mathcal{P}$ are chosen as follows: Let $e_1, \dots, e_d$ denote the standard unit vectors in $\mathbb{R}^d$. For now, assume that $d = n/\lg(1/\varepsilon)$ and $\varepsilon \in (\lg^{0.5001} n/\sqrt{d}, 1)$. We will later show how to generalize the proof to the full range of d. For any set $S \subset [d]$ of $k = \varepsilon^{-2}/256$ indices, define a vector $y_S := \sum_{j \in S} e_j/\sqrt{k}$. A vector y_S has the property that $\langle y_S, e_j \rangle = 0$ if $j \notin S$ and $\langle y_S, e_j \rangle = 16\varepsilon$ if $j \in S$. The crucial property here is that there is a gap of 16ε between the inner products depending on whether or not $j \in S$. Now if f is a mapping to $\mathbb{R}^m$ that satisfies the JL-property (1) for $P = \{0, e_1, \dots, e_d, y_S\}$, then first off, we can assume $f(0) = 0$ since pairwise distances are translation invariant. From this it follows that f must preserve norms of the vectors $x \in P$ to within $(1+\varepsilon)$ since

$$\begin{aligned}(1-\varepsilon)\|x\|_2^2 &= (1-\varepsilon)\|x-0\|_2^2 \le \|f(x)-f(0)\|_2^2 \\ &= \|f(x)\|_2^2 = \|f(x)-f(0)\|_2^2 \\ &\le (1+\varepsilon)\|x-0\|_2^2 \\ &= (1+\varepsilon)\|x\|_2^2.\end{aligned}$$

We then have that f must preserve inner products $\langle e_j, y_S \rangle$ up to an additive of 4ε. This can be seen by the following calculations, where $v \pm X$ denotes the interval $[v - X, v + X]$:

$$\begin{aligned}\|f(e_j) - f(y_S)\|_2^2 &= \|f(e_j)\|_2^2 + \|f(y_S)\|_2^2 \\ &\quad - 2\langle f(e_j), f(y_S)\rangle \Rightarrow \\ 2\langle f(e_j), f(y_S)\rangle &\in (1\pm\varepsilon)\|e_j\|_2^2 + (1\pm\varepsilon)\|y_S\|_2^2 \\ &\quad - (1\pm\varepsilon)\|e_j - y_S\|_2^2 \Rightarrow \\ 2\langle f(e_j), f(y_S)\rangle &\in 2\langle e_j, y_S\rangle \pm \varepsilon(\|e_j\|_2^2 + \|y_S\|_2^2 \\ &\quad + \|e_j - y_S\|_2^2) \Rightarrow \\ \langle f(e_j), f(y_S)\rangle &\in \langle e_j, y_S\rangle \pm 4\varepsilon.\end{aligned}$$

This means that after applying f, there remains a gap of $(16-8)\varepsilon = 8\varepsilon$ between $\langle f(e_j), f(y_S)\rangle$ depending on whether or not $j \in S$. With this observation, we are ready to describe the point sets in $\mathcal{P}$ (in fact they will not be point sets, but rather ordered sequences of points, possibly with repetition). Let $Q = n - d - 1$. For every choice of Q sets $S_1, \dots, S_Q \subset [d]$ of k indices each, we add a point set P to $\mathcal{P}$. The sequence P is simply $(0, e_1, \dots, e_d, y_{S_1}, \dots, y_{S_Q})$. This gives us a family $\mathcal{P}$ of size $\binom{d}{k}^Q$. If we look at JL embeddings for all of these point sets $f_1(P_1), f_2(P_2), \dots,$ then intuitively these embeddings have to be quite different. This is true since $f_i(P_i)$ uniquely determines P_i simply by computing all inner products between the $f_i(e_j)$'s and $f_i(y_{S_\ell})$'s. The problem we now face is that there are infinitely many sets of n points in $\mathbb{R}^m$ that one can embed to. We thus need to discretize $\mathbb{R}^m$ in a careful manner and argue that there are not enough n-sized sets of points in this discretization to uniquely embed each P_i when m is too low.

Encoding Argument To give a formal proof that there are not enough ways to embed the point sets in $\mathcal{P}$ into $\mathbb{R}^m$ when m is low, we give an encoding argument. More specifically, we assume that it is possible to embed every point set in $\mathcal{P}$ into $\mathbb{R}^m$ while preserving pairwise distances to within $(1+\varepsilon)$. We then present an algorithm that based on this assumption can take any point set $P \in \mathcal{P}$ and encode it into a bit string of length $O(nm)$. The encoding guarantees that P can be uniquely recovered from the encoding. The encoding algorithm thus effectively defines an injective mapping g from $\mathcal{P}$ to $\{0,1\}^{O(nm)}$. Since g is injective, we must have $|\mathcal{P}| \le 2^{O(nm)}$. But $|\mathcal{P}| = \binom{d}{k}^Q = (\varepsilon^2 n/\lg(1/\varepsilon))^{\Omega(\varepsilon^{-2} n)}$ and we can conclude $m = \Omega(\varepsilon^{-2} \lg(\varepsilon^2 n/\lg(1/\varepsilon)))$. For $\varepsilon > 1/n^{0.4999}$, this is $m = \Omega(\varepsilon^{-2} \lg n)$.

$$A\left\{\begin{array}{|c|}\hline \hat{f}(e_1)^T\\ \hline \hat{f}(e_2)^T\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \hat{f}(e_d)^T\\ \hline\end{array}\right. \cdot \begin{array}{|c|}\hline \\ \hat{f}_i(y_{S_\ell})\\ \\ \hline\end{array} = \left.\begin{array}{|c|}\hline \left\langle \hat{f}(e_1), \hat{f}(y_{S_\ell}) \right\rangle\\ \hline \left\langle \hat{f}(e_2), \hat{f}(y_{S_\ell}) \right\rangle\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \cdots\\ \hline \left\langle \hat{f}(e_d), \hat{f}(y_{S_\ell}) \right\rangle\\ \hline\end{array}\right\} v_\ell$$

Figure 1: Notation to describe a more efficient encoding of $P \in \mathcal{P}$.

First Attempt The difficult part is to design an encoding algorithm that yields an encoding of size $O(nm)$ bits. A natural first attempt would go as follows: recall that any JL-embedding f for a point set $P \in \mathcal{P}$ (where f may depend on P) must preserve gaps in $\langle f(e_j), f(y_{S_\ell})\rangle$'s depending on whether or not $j \in S_\ell$. This follows simply by preserving distances to within a factor $(1+\varepsilon)$ as argued above. If we can give an encoding that allows us to recover approximations $\hat{f}(e_j)$ of $f(e_j)$ and $\hat{f}(y_{S_\ell})$ of $f(y_{S_\ell})$ such that $\|\hat{f}(e_j) - f(e_j)\|_2^2 \leq \varepsilon$ and $\|\hat{f}(y_{S_\ell}) - f(y_{S_\ell})\|_2^2 \leq \varepsilon$, then by the triangle inequality, the distance $\|\hat{f}(e_j) - \hat{f}(y_{S_\ell})\|_2^2$ is also a $(1+O(\varepsilon))$ approximation to $\|e_j - y_{S_\ell}\|_2^2$ and the gap between inner products would be preserved. To encode sufficiently good approximations $\hat{f}(e_j)$ and $\hat{f}(y_{S_\ell})$, one could do as follows: since norms are roughly preserved by f, we must have $\|f(e_j)\|_2^2, \|f(y_{S_\ell})\|_2^2 \leq 1+\varepsilon$. Letting B_2^m denote the ℓ_2 unit ball in $\mathbb{R}^m$, we could choose some fixed covering C_2 of $(1+\varepsilon)B_2^m$ with translated copies of εB_2^m. Since $f(e_j), f(y_{S_\ell}) \in (1+\varepsilon)B_2^m$, we can find translations $c_2(f(e_j)) + \varepsilon B_2^m$ and $c_2(f(y_{S_\ell})) + \varepsilon B_2^m$ of εB_2^m in C_2, such that these balls contain $f(e_j)$ and $f(y_{S_\ell})$ respectively. Letting $\hat{f}(e_j) = c_2(f(e_j))$ and $\hat{f}(y_{S_\ell}) = c_2(f(y_{S_\ell}))$ be the centers of these balls, we can encode an approximation of $f(e_j)$ and $f(y_{S_\ell})$ using $\lg|C_2|$ bits by specifying indices into C_2. Unfortunately, covering $(1+\varepsilon)B_2^m$ by εB_2^m needs $|C_2| = 2^{\Omega(m\lg(1/\varepsilon))}$ since the volume ratio between $(1+\varepsilon)B_2^m$ and εB_2^m is $(1/\varepsilon)^{\Omega(m)}$. The $\lg(1/\varepsilon)$ factor loss leaves us with a lower bound on m of no more than $m = \Omega(\varepsilon^{-2}\lg(\varepsilon^2 n/\lg(1/\varepsilon))/\lg(1/\varepsilon))$, roughly recovering the lower bound of Alon [Alo03] by a different argument.

Full Proof The key idea to reduce the length of the encoding to $O(nm)$ is as follows: First observe that we chose $d = n/\lg(1/\varepsilon)$. Thus we can spend up to $O(m\lg(1/\varepsilon))$ bits encoding each $f(e_j)$'s. Thus we simply encode approximations $\hat{f}(e_j)$ by specifying indices into a covering C_2 of $(1+\varepsilon)B_2^m$ by εB_2^m as outlined above.

For the $f(y_{S_\ell})$'s, we have to be more careful as we cannot afford $m\lg(1/\varepsilon)$ bits for each. First, we define the $d \times m$ matrix A having the $\hat{f}(e_j) = c_2(f(e_j))$ as rows (see Figure 1). Note that this matrix can be reconstructed from the part of the encoding specifying the $\hat{f}(e_j)$s. Now observe that the j'th coordinate of $v_\ell = Af(y_{S_\ell})$ is equal to $\langle \hat{f}(e_j), f(y_{S_\ell})\rangle$. This is within $O(\varepsilon)$ of $\langle e_j, y_{S_\ell}\rangle$. The coordinates of v_ℓ thus determine S_ℓ due to the gap in inner products depending on whether $j \in S_\ell$ or not. We therefore seek to encode the v_ℓ efficiently. Since the v_ℓ are in $\mathbb{R}^d$, this seems quite hopeless to do in $O(m)$ bits per v_ℓ. The key observation is that they lie in an m-dimensional subspace of $\mathbb{R}^d$, namely in the column space of A. This observation will allow us to get down to just $O(m)$ bits. We are ready to give the remaining details.

Let W denote the subspace of $\mathbb{R}^d$ spanned by the columns of A. We have $\dim(W) \leq m$. Define T as the convex body

$$T := B_\infty^d \cap W.$$

That is, T is the intersection of the subspace W with the d-dimensional ℓ_∞ unit ball B_∞^d. Now let C_∞ be a minimum cardinality covering of $(22\varepsilon)T$ by translated copies of εT, computed by any deterministic procedure that depends only on T. Since T is origin symmetric, by Corollary 1 it follows that $|C_\infty| \leq 2^{m \lg 45}$. To encode the vectors $y_{S_1}, \dots, y_{S_Q}$ we make use of the following lemma, whose proof we give in Section 3.1:

Lemma 3. *For every e_j and y_{S_ℓ} in P, we have*

$$|\langle \hat{f}(e_j), f(y_{S_\ell})\rangle - \langle e_j, y_{S_\ell}\rangle| \leq 6\varepsilon.$$

From Lemma 3, it follows that $|\langle \hat{f}(e_j), f(y_{S_\ell})\rangle| \leq 6\varepsilon + \langle e_j, y_{S_\ell}\rangle \leq 22\varepsilon$ for every e_j and y_{S_ℓ} in P. Since the j'th coordinate of $Af(y_{S_\ell})$ equals $\langle \hat{f}(e_j), f(y_{S_\ell})\rangle$, it follows that $Af(y_{S_\ell}) \in (22\varepsilon)T$. Using this fact, we encode each y_{S_ℓ} by finding some vector $c_\infty(y_{S_\ell})$ such that $c_\infty(y_{S_\ell}) + \varepsilon T$ is a convex shape in the covering C_∞ and $Af(y_{S_\ell}) \in c_\infty(y_{S_\ell}) + \varepsilon T$. We write down $c_\infty(y_{S_\ell})$ as an index into C_∞. This costs a total of $Qm \lg 45 = O(Qm)$ bits over all y_{S_ℓ}. We now describe our decoding algorithm.

Decoding Algorithm To recover $P = \{0, e_1, \dots, e_d, y_{S_1}, \dots, y_{S_Q}\}$ from the above encoding, we only have to recover $y_{S_1}, \dots, y_{S_Q}$ as $\{0, e_1, \dots, e_d\}$ is the same for all $P \in \mathcal{P}$. We first reconstruct the matrix A. We can do this since C_2 was chosen independently of P and thus by the indices encoded into C_2, we recover $c_2(e_j) = \hat{f}(e_j)$ for $j = 1, \dots, d$. These are the rows of A. Then given A, we know T. Knowing T, we compute C_∞ since it was constructed via a deterministic procedure depending only on T. This finally allows us to recover $c_\infty(y_{S_1}), \dots, c_\infty(y_{S_Q})$. What remains is to recover $y_{S_1}, \dots, y_{S_Q}$. Since y_{S_ℓ} is uniquely determined from the set $S_\ell \subseteq \{1, \dots, d\}$ of k indices, we focus on recovering this set of indices for each y_{S_ℓ}.

For $\ell = 1, \dots, Q$ recall that $Af(y_{S_\ell})$ is in $c_\infty(y_{S_\ell}) + \varepsilon T$. Observe now that:

$$\begin{aligned} Af(y_{S_\ell}) \in c_\infty(y_{S_\ell}) + \varepsilon T &\quad \Rightarrow \\ Af(y_{S_\ell}) - c_\infty(y_{S_\ell}) \in \varepsilon T &\quad \Rightarrow \\ \|Af(y_{S_\ell}) - c_\infty(y_{S_\ell})\|_\infty \leq \varepsilon. & \end{aligned}$$

But the j'th coordinate of $Af(y_{S_\ell})$ is $\langle \hat{f}(e_j), f(y_{S_\ell})\rangle$. We combine the above with Lemma 3 to deduce $|(c_\infty(y_{S_\ell}))_j - \langle e_j, y_{S_\ell}\rangle| \leq 7\varepsilon$ for all j. We thus have that $(c_\infty(y_{S_\ell}))_j \leq 7\varepsilon$ for $j \notin S_i$ and $(c_\infty(y_{S_\ell}))_j \geq 9\varepsilon$ for $j \in S_\ell$. We finally conclude that the set S_ℓ, and thus y_{S_ℓ}, is uniquely determined from $c_\infty(y_{S_\ell})$.

Analysis We finally analyse the size of the encoding produced by the above procedure and derive a lower bound on m. Recall that the encoding procedure produces a total of $dm \lg(1 + 4/\varepsilon) + O(Qm) = O(nm)$ bits. But $|\mathcal{P}| \geq \left(\binom{d}{k}/2\right)^Q \geq (d/(2k))^{kQ} = (d/(2k))^{k(n-d-1)} \geq (d/(2k))^{kn/2}$. We therefore must have

$$\begin{aligned} nm &= \Omega(kn \lg(d/k)) \Rightarrow \\ m &= \Omega(\varepsilon^{-2} \lg(\varepsilon^2 n / \lg(1/\varepsilon))). \end{aligned}$$

Since we assume $\varepsilon > \lg^{0.5001} n/\sqrt{d} \geq \lg^{0.5001} n/\sqrt{n}$, this can be simplified to

$$m = \Omega(\varepsilon^{-2} \lg(\varepsilon^2 n)).$$

This shows that $m = \Omega(\varepsilon^{-2} \lg(\varepsilon^2 n))$ for $d = n/\lg(1/\varepsilon)$ and $\varepsilon \in (\lg^{0.5001} n/\sqrt{d}, 1)$. The following paragraph shows how to handle the remaining values of d.

Handling Other Values of d For $d > n/\lg(1/\varepsilon)$, the proof is easy: Simply repeat the above construction using only the first $n/\lg(1/\varepsilon)$ standard unit vectors in the point sets of $\mathcal{P}$. This reproves the above lower bound, with the only further restriction that $\varepsilon \in (\lg^{0.5001} n/\sqrt{\min\{d,n\}}, 1)$ as opposed to $\varepsilon \in (\lg^{0.5001} n/\sqrt{d}, 1)$.

For $d < n/\lg(1/\varepsilon)$ and $\varepsilon \in (\lg^{0.5001} n/\sqrt{d}, 1)$, assume for the sake of contradiction that it is possible to embed into $o(\varepsilon^{-2} \lg(\varepsilon^2 n))$ dimensions. Now take any point set P in $\mathbb{R}^{d'}$ with $d' = n/\lg(1/\varepsilon)$ and apply a JL transform into d dimensions on it, obtaining a point set P' in d dimensions. This new point set has all distances preserved to within $(1 + O(\sqrt{\lg n/d}))$ (by the standard JL upper bound). Next apply the hypothetical JL transform in d dimensions to reduce the target dimension to $o(\varepsilon^{-2}\lg(\varepsilon^2 n))$. Distances are now preserved to within $(1 + O(\sqrt{\lg n/d}))(1 + \varepsilon)$. Since we assumed $\varepsilon > \lg^{0.5001} n/\sqrt{d}$, we have that $(1 + O(\sqrt{\lg n/d})) = (1 + o(\varepsilon))$, which implies $(1 + O(\sqrt{\lg n/d}))(1 + \varepsilon) = (1 + O(\varepsilon))$. This contradicts the lower bound for $d' = n/\lg(1/\varepsilon)$ dimensions.

3.1 Proof of Lemma 3

In this section, we prove the lemma:

Restatement of Lemma 3. *For every e_j and y_{S_ℓ} in P, we have*

$$|\langle \hat{f}(e_j), f(y_{S_\ell})\rangle - \langle e_j, y_{S_\ell}\rangle| \leq 6\varepsilon.$$

Proof. First note that:

$$\begin{aligned}
\langle \hat{f}(e_j), f(y_{S_\ell})\rangle &= \\
\langle c_2(e_j) - f(e_j) + f(e_j), f(y_{S_\ell})\rangle &= \\
\langle f(e_j), f(y_{S_\ell})\rangle + \langle c_2(e_j) - f(e_j), f(y_{S_\ell})\rangle &\in \\
\langle f(e_j), f(y_{S_\ell})\rangle \pm \|c_2(e_j) - f(e_j)\|_2 \|f(y_{S_\ell})\|_2.
\end{aligned}$$

Since C_2 was a covering with εB_2^m, we have $\|c_2(e_j) - f(e_j)\|_2 \leq \varepsilon$. Recall that $\|f(y_{S_\ell})\|_2^2 \leq (1+\varepsilon)$. This in particular implies that $\|f(y_{S_\ell})\|_2 \leq 2$. We thus have:

$$\langle \hat{f}(e_j), f(y_{S_\ell}))\rangle \quad \in \quad \langle f(e_j), f(y_{S_\ell})\rangle \pm 2\varepsilon. \tag{4}$$

To bound $\langle f(e_j), f(y_{S_\ell})\rangle$, observe that

$$\begin{aligned}
\|f(e_j) - f(y_{S_\ell})\|_2^2 &= \\
\|f(e_j)\|_2^2 + \|f(y_{S_\ell})\|_2^2 - 2\langle f(e_j), f(y_{S_\ell})\rangle.
\end{aligned}$$

This implies that

$$\begin{aligned}
2\langle f(e_j), f(y_{S_\ell})\rangle &\in \\
\|e_j\|_2^2(1\pm\varepsilon) + \|y_{S_\ell}\|_2^2(1\pm\varepsilon) - \|e_i - y_{S_\ell}\|_2^2(1\pm\varepsilon) &\subseteq \\
2\langle e_j, y_{S_\ell}\rangle \pm \varepsilon(\|e_j\|_2^2 + \|y_{S_\ell}\|_2^2 + \|e_j - y_{S_\ell}\|_2^2) &\subseteq \\
2\langle e_j, y_{S_\ell}\rangle \pm \varepsilon(4(\|e_j\|_2^2 + \|y_{S_\ell}\|_2^2))
\end{aligned}$$

That is, we have

$$\langle f(e_j), f(y_{S_\ell})\rangle \quad \in \quad \langle e_j, y_{S_\ell}\rangle \pm 2\varepsilon(\|e_j\|_2^2 + \|y_{S_\ell}\|_2^2)$$

Both the e_j's and y_{S_ℓ}'s have unit norm, hence

$$\langle f(e_j), f(y_{S_\ell})\rangle \quad \in \quad \langle e_j, y_{S_\ell}\rangle \pm 4\varepsilon$$

Inserting this in (4), we obtain

$$\langle \hat{f}(e_j), f(y_{S_\ell})\rangle \quad \in \quad \langle e_j, y_{S_\ell}\rangle \pm 6\varepsilon.$$

□

Acknowledgments

We thank Oded Regev for pointing out a simplification to our initial argument for handling the case $d < n/\lg(1/\varepsilon)$, and for his permission to include the simpler argument here.

K.G.L. is supported by Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation, grant DNRF84, a Villum Young Investigator Grant and an AUFF Starting Grant. J.N. did this work while supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

References

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple construction of almost k-wise independent random variables. *Random Struct. Algorithms*, 3(3):289–304, 1992.

[AK17] Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *Proceedings of the 58th Annual Symposium on Foundations of Computer Science (FOCS)*, 2017.

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

[BZMD15] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, 2015.

[CEM+15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Mădălina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the 47th ACM Symposium on Theory of Computing (STOC)*, 2015. Full version at `http://arxiv.org/abs/1410.6801v3`.

[CRT06] Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

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

[HIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

[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.

[KOR00] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30(2):457–474, 2000.

[LN16] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction. In *Proceedings of the 43rd International Colloquium on Automata, Languages and Programming (ICALP)*, 2016.

[Lev83] Vladimir I. Levenshtein. Bounds for packings of metric spaces and some of their applications (in Russian). *Problemy Tekhn. Kibernet. Robot.* vol. 40, pages 43–110, 1983.

[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.

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

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

[Tho96] Anthony C. Thompson. *Minkowski Geometry*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 1996.

[Wel74] Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20, May 1974.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.