Optimality of the Johnson-Lindenstrauss Lemma

Kasper Green Larsen* Jelani Nelson†

September 7, 2016

Abstract

For any integers $d, n \geq 2$ 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, 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 proved that for ε smaller than some universal constant ε_0, 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 by Alon [Alo03], who showed the existence of an n point set $X \subset \mathbb{R}^n$, such that any f providing (1) must have $m = \Omega(\min\{n, \varepsilon^{-2} \lg n / \lg(1/\varepsilon)\})$. 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 $c > 0$. This bound falls short of the JL lemma for any $\varepsilon = o(1)$.

*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*

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

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 the requirement $\varepsilon > (\lg^{0.5001} n)/\sqrt{\min\{n,d\}}$ is necessary for such a lower bound to apply, up to the $\lg^{0.5001} n$ factor.

It is worth mentioning that the arguments in previous work [Wel74, Alo03, LN16] all produced hard point sets P which were *incoherent*: every $x \in P$ had unit ℓ_2 norm, and $\forall x \neq y \in P$ one had $|\langle x, y\rangle| = O(\varepsilon)$ (to be more precise, the argument in [LN16] produced a random point set which was hard to embed into low dimension with high probability, but was also incoherent with high probability; the arguments in [Wel74, Alo03] necessarily required P to be incoherent). Unfortunately though, it is known that for $\varepsilon < 2^{-\omega(\sqrt{\lg n})}$ an embedding with $m = o(\varepsilon^{-2} \lg n)$ satisfying (1) exists, beating the guarantee of the JL lemma. The construction is based on Reed-Solomon codes (see for example [NNW14]). Thus proving Theorem 2 requires a very different construction of a hard point set when compared with previous work.

1.1 Related Results

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 a fixed 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*

$$\forall u \in \mathbb{R}^d, \underset{\Pi \sim \mathcal{D}_{\varepsilon,\delta}}{\mathbb{P}}(|\|\Pi u\|_2 - \|u\|_2| > \varepsilon \|u\|_2) < \delta. \tag{2}$$

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 (2) 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 is optimal.

2 Proof Overview

In the following, we give a high level introduction of the main ideas in our 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 $\varepsilon \in (0,1)$. For any set $S \subset [d]$ of $k = \varepsilon^{-2}/c_0^2$ indices, define a vector $y_S := \sum_{j \in S} e_j/\sqrt{k}$. Here c_0 is a sufficiently large constant. A vector y_S has the property that $\langle y_S, e_j \rangle = 0$ if $j \notin S$ and $\langle y_S, e_j \rangle = c_0\varepsilon$ if $j \in S$. The crucial property here is that there is a gap of $c_0\varepsilon$ 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

$$(1-\varepsilon)\|x\|_2^2 = (1-\varepsilon)\|x-0\|_2^2 \le \|f(x)-f(0)\|_2^2 = \|f(x)\|_2^2 = \|f(x)-f(0)\|_2^2 \le (1+\varepsilon)\|x-0\|_2^2 = (1+\varepsilon)\|x\|_2^2.$$

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

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

This means that after applying f, there remains a gap of $(c_0 - 8)\varepsilon = \Omega(\varepsilon)$ 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 $\mathcal{P}$. 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 point set 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 $P_i \in \mathcal{P}$ and encode it into a bit string of length $O(nm)$. The encoding guarantees that P_i 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 $|\mathcal{P}| = \binom{d}{k}^Q \ge (\varepsilon^2 n/\lg(1/\varepsilon))^{(\varepsilon^{-2}(n - n/\lg(1/\varepsilon) - 1)/c_0^2)}$ and we can conclude $m = \Omega(\varepsilon^{-2}\lg(\varepsilon^2 n/\lg(1/\varepsilon)))$. For $\varepsilon > 1/n^{0.4999}$, this is $m = \Omega(\varepsilon^{-2}\lg n)$.

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_i for a point set $P_i \in \mathcal{P}$ must preserve gaps in $\langle f_i(e_j), f_i(y_{S_\ell})\rangle$'s depending on whether or not $j \in S_\ell$. This follows simply by preserving distances to within a factor $(1+\varepsilon)$. If we can give an encoding that allows us to recover approximations $\hat{f}_i(e_j)$ of $f_i(e_j)$ and $\hat{f}_i(y_{S_\ell})$ of $f_i(y_{S_\ell})$ such that $\|\hat{f}_i(e_j) - f_i(e_j)\|_2^2 \le \varepsilon$ and $\|\hat{f}_i(y_{S_\ell}) - f_i(y_{S_\ell})\|_2^2 \le \varepsilon$, then by the triangle inequality, the distance $\|\hat{f}_i(e_j) - \hat{f}_i(y_{S_\ell})\|_2^2$ 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 $\hat{f}_i(e_j)$ and $\hat{f}_i(y_{S_\ell})$, one could do as follows: Since norms are roughly preserved by f_i, we must have $\|f_i(e_j)\|_2^2, \|f_i(y_{S_\ell})\|_2^2 \leq 1+\varepsilon$. 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 $f_i(e_j), f_i(y_{S_\ell}) \in (1+\varepsilon)B_2^m$, we can find translations $c_2(f_i(e_j)) + \varepsilon B_2^m$ and $c_2(f_i(y_{S_\ell})) + \varepsilon B_2^m$ of εB_2^m in C_2, such that these balls contain $f_i(e_j)$ and $f_i(y_{S_\ell})$ respectively. Letting $\hat{f}_i(e_j) = c_2(f_i(e_j))$ and $\hat{f}_i(y_{S_\ell}) = c_2(f_i(y_{S_\ell}))$ be the centers of these balls, we can encode an approximation of $f_i(e_j)$ and $f_i(y_{S_\ell})$ 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.

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 $f_i(e_j)$'s. Thus we simply encode approximations $\hat{f}_i(e_j)$ by specifying indices into a covering C_2 of $(1+\varepsilon)B_2^m$ by εB_2^m as outlined above. For the $f_i(y_{S_\ell})$'s, we have to be more careful. First, we define the $d \times m$ matrix A having the $\hat{f}_i(e_j)$ as rows. Note that this matrix can be reconstructed from the part of the encoding specifying the $\hat{f}_i(e_j)$s. Now observe that the j'th coordinate of $Af_i(y_{S_\ell})$ is within $O(\varepsilon)$ of $\langle e_j, y_{S_\ell}\rangle$. The coordinates of $Af_i(y_{S_\ell})$ thus determine S_ℓ. We therefore seek to encode $Af_i(y_{S_\ell})$ efficiently.

To encode $Af_i(y_{S_\ell})$, first note that $\|Af_i(y_{S_\ell})\|_\infty = O(\varepsilon)$. If W denotes the $\leq m$-dimensional subspace spanned by the columns of A, we also have that $Af_i(y_{S_\ell}) \in W$. Now define the convex body $T := B_\infty^d \cap W$, where B_∞^d denotes the ℓ_∞ unit cube in $\mathbb{R}^d$. Then $Af_i(y_{S_\ell}) \in O(\varepsilon)\cdot T$. Now recall that there is a gap of $\Omega(\varepsilon)$ between inner products $\langle \hat{f}_i(e_j), f_i(y_{S_\ell})\rangle$ depending on whether $j \in S_\ell$ or not. Letting c_1 be a constant such that the gap is more than $2c_1\varepsilon$, this implies that if we approximate $Af_i(y_{S_\ell})$ by a point $\hat{f}_i(y_{S_\ell})$ such that $(\hat{f}_i(y_{S_\ell}) - Af_i(y_{S_\ell})) \in c_1\varepsilon \cdot B_\infty^d$, then the coordinates of $\hat{f}_i(y_{S_\ell})$ still uniquely determine the indices $j \in S_\ell$. Exploiting that $Af_i(y_{S_\ell}) \in O(\varepsilon)\cdot T$, we therefore create a covering C_∞ of $O(\varepsilon)\cdot T$ by translated copies of $c_1\varepsilon \cdot T$ and approximate $Af_i(y_{S_\ell})$ by a convex body in C_∞ containing it. The crucial property of this construction is that the volume ratio between $O(\varepsilon)\cdot T$ and $c_1\varepsilon\cdot T$ is only $2^{O(m)}$ and we can have $|C_\infty| = 2^{O(m)}$. Specifying indices into C_∞ thus costs only $O(m)$ bits and we have obtained the desired encoding algorithm.

Handling Small d. In the proof sketch above, we assumed $d = n/\lg(1/\varepsilon)$. As d went into $|\mathcal{P}|$, the above argument falls apart when d is much smaller than n and would only yield a lower bound of $m = \Omega(\varepsilon^{-2}\lg(\varepsilon^2 d/\lg(1/\varepsilon)))$. Fortunately, we do not have to use completely orthogonal vectors $e_1, \ldots, e_d$ in our construction. What we really need is that we have a base set of vectors $x_1, \ldots, x_\Delta$ and many subsets S of k indices in $[\Delta]$, such that there is a gap of $\Omega(\varepsilon)$ between inner products $\langle x_j, \sum_{i\in S} x_i/\sqrt{k}\rangle$ depending on whether or not $j \in S$. To construct such $x_1, \ldots, x_\Delta$ when d is small, we argue probabilistically: We pick $x_1, \ldots, x_\Delta$ as uniform random standard gaussians scaled by a factor $1/\sqrt{d}$, i.e. each coordinate of each x_i is independently $\mathcal{N}(0, 1/d)$ distributed. We show that with non-zero probability, at least half of all k-sized sets S of indices in $[\Delta]$ have the desired property of sufficiently large gaps in the inner products. The proof of this is given in Section 4. Here we also want to remark that it would be more natural to require that *all* k-sized sets of indices S have the gap property. Unfortunately, this requires us to union bound over $\binom{\Delta}{k}$ different subsets and we would obtain a weaker lower bound for small d.

3 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. $\square$

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

4 Nearly Orthogonal Vectors

In the proof overview in Section 2, we argued towards the end that we need to show the existence of a large set of base vectors $x_1, \ldots, x_\Delta$ and many k-sized sets of indices $S \subset [\Delta]$, such that there is a gap between $\langle x_i, \sum_{j \in S} x_j/\sqrt{k}\rangle$ depending on whether or not $i \in S$. Below we formally define such collections of base vectors and state two lemmas regarding how large sets we can construct and what properties such sets have. We defer the proofs of the lemmas to Section 6.

Definition 1. *Let $X = \{x_1, \ldots, x_N\}$ be a set of vectors in $\mathbb{R}^d$ and $\mathcal{F}$ a collection of k-sized subsets of $[N]$. For any $0 < \mu < 1$ and integer $k \geq 1$, we say that $(X, \mathcal{F})$ is k-wise μ-incoherent if for every vector $x_j \in X$, the following holds:*

1. *$|\|x_j\|_2^2 - 1| \leq \mu$.*
2. *For every $S \in \mathcal{F}$, it holds that $|\langle x_j, \sum_{i \in S: i \neq j} x_i/\sqrt{k}\rangle| \leq \mu$.*

Our first step is to show the existence of a large $(X, \mathcal{F})$ that is k-wise μ-incoherent:

Lemma 3. *For any $0 < \mu < 1$, $1 \leq N \leq \max\{d, e^{O(\mu^2 d)}\}$ and integer $1 \leq k \leq N$, there exists $(X, \mathcal{F})$ with $X \subset \mathbb{R}^d$ and $|X| = N$, such that $(X, \mathcal{F})$ is k-wise μ-incoherent and*

$$|\mathcal{F}| \geq \binom{N}{k}/2.$$

This lemma is proved in Section 6.

The following property of k-wise μ-incoherent pairs $(X, \mathcal{F})$ plays a crucial role in our lower bound proof:

Lemma 4. *Let $(X, \mathcal{F})$ be k-wise μ-incoherent for some $0 < \mu < 1$ and $k \geq 1$. Let $S \in \mathcal{F}$ and define $y = \sum_{i \in S} x_i/\sqrt{k}$. Then y satisfies:*

1. *$\|y\|_2^2 \leq 1 + (\sqrt{k} + 1)\mu$.*
2. *For a vector $x_j \in X$ such that $j \notin S$, we have $|\langle y, x_j\rangle| \leq \mu$.*
3. *For a vector $x_j \in X$ such that $j \in S$, we have $(1 - \mu)/\sqrt{k} - \mu \leq |\langle y, x_j\rangle| \leq (1 + \mu)/\sqrt{k} + \mu$.*

The proof of this lemma can also be found in Section 6.

5 Lower Bound Proof

The goal of this section is to prove Theorem 2:

Restatement of 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*

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

We assume throughout the proof that $\varepsilon \in (\lg^{0.5001} n/\sqrt{\min\{n,d\}}, 1)$ as required in the theorem.

We now describe a collection of point sets $\mathcal{P} = \{P_1, \dots\}$ for which at least one point set $P_i \in \mathcal{P}$ cannot be embedded into $o(\varepsilon^{-2}\lg(\varepsilon^2 n))$ dimensions while preserving all pairwise distances to within a factor $(1 \pm \varepsilon)$.

Let $\mu = \varepsilon/2$, $k = (\varepsilon^{-2}/2^{20})$ and $\Delta = n/\lg(1/\varepsilon)$. For this choice of μ, k and Δ, and using $\varepsilon > \lg^{0.5001} n/\sqrt{\min\{n,d\}}$, we have that $k < \Delta/2$ and $e^{\Omega(\mu^2 d)} = \omega(n)$. Hence it follows from Lemma 3 that there exists $(X, \mathcal{F})$ that is k-wise μ-incoherent with $|X| = \Delta$ and $|\mathcal{F}| \geq \binom{\Delta}{k}/2$. For every sequence of $Q = n - \Delta - 1$ sets $S_1, \dots, S_Q \in \mathcal{F}$, we add a point set P to $\mathcal{P}$. For sets $S_1, \dots, S_Q$ we construct P as follows: First we add the zero-vector and X to P. Let $x_1, \dots, x_\Delta$ denote the vectors in X. The remaining Q vectors are denoted $y_1, \dots, y_Q$ and are defined as

$$y_i = \sum_{j \in S_i} x_j/\sqrt{k}.$$

The point set P is thus the ordered sequence of points $\{0, x_1, \dots, x_\Delta, y_1, \dots, y_Q\}$. This concludes the description of the hard point sets. Observe that $|\mathcal{P}| \geq \left(\binom{\Delta}{k}/2\right)^Q$. Also, Lemma 4 implies the following for our choice of k and μ:

Corollary 2. *Let $(X, \mathcal{F})$ be $(\varepsilon^{-2}/2^{20})$-wise $(\varepsilon/2)$-incoherent. Let $S \in \mathcal{F}$ and define $y = \sum_{i \in S} x_i/\sqrt{k}$. Then y satisfies:*

1. $\|y\|_2^2 \leq 2$.
2. *For a vector $x_j \in X$ such that $j \notin S$, we have $|\langle y, x_j\rangle| \leq \varepsilon/2$.*
3. *For a vector $x_j \in X$ such that $j \in S$, we have $2^8\varepsilon \leq |\langle y, x_j\rangle| \leq 2^{11}\varepsilon$.*

Proof. The first item follows since $1 + (\sqrt{k}+1)\mu \leq 1 + 1/2^{11} + \varepsilon/2 < 2$. The second item follows by choice of μ and item 2 from Lemma 4. For the third item: $(1-\mu)/\sqrt{k} - \mu \geq 2^{10}\varepsilon - 2^9\varepsilon^2 - \varepsilon/2 \geq 2^8\varepsilon$ and $(1+\mu)/\sqrt{k} + \mu \leq 2^{10}\varepsilon + 2^9\varepsilon^2 + \varepsilon/2 \leq 2^{11}\varepsilon$. $\square$

Our lower bound follows via an encoding argument. More specifically, we assume that for every set $P \subset \mathbb{R}^d$ of n points, there exists an embedding $f_P : P \to \mathbb{R}^m$ satisfying:

$$\forall x, y \in P : (1-\varepsilon)\|x-y\|_2^2 \leq \|f_P(x) - f_P(y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2 \tag{3}$$

Under this assumption, we show how to encode and decode a set of vectors $P_i \in \mathcal{P}$ using $O(nm)$ bits. Since $|\mathcal{P}| \geq \left(\binom{\Delta}{k}/2\right)^Q$, any encoding that allows unique recovery of the vectors sets in $\mathcal{P}$ must necessarily use $\lg|\mathcal{P}| = Q(\lg\binom{\Delta}{k} - 1) \geq Q(k\lg(\Delta/k) - 1)$ bits on at least one point set $P_i \in \mathcal{P}$. This will establish a lower bound on m.

Encoding Algorithm. First, the encoder and decoder agree on a covering C_2 of $2B_2^m$ by translated copies of εB_2^m. Lemma 2 guarantees that there exists such a covering C_2 with $|C_2| \leq 2^{m\lg(1+4/\varepsilon)}$.

Now, given a set of vectors/points $P \in \mathcal{P}$, where $P = \{0, x_1, \dots, x_\Delta, y_1, \dots, y_Q\}$, the encoder first applies f_P to all points in P. Henceforth we abbreviate f_P as f. Since pairwise distances are invariant under translation, we assume wlog. that $f(0) = 0$. Since f satisfies (3), we must have

$$\begin{aligned} \forall x_i \in P : \|f(x_i) - f(0)\|_2^2 \leq (1+\varepsilon)\|x_i - 0\|_2^2 \quad &\Rightarrow \\ \forall x_i \in P : \|f(x_i)\|_2^2 \leq (1+\varepsilon)\|x_i\|_2^2 \leq (1+\varepsilon)(1+\varepsilon/2) \leq 4 \quad &\Rightarrow \\ \forall x_i \in P : \|f(x_i)\|_2 \leq 2. \end{aligned}$$

Take each $x_i \in P$ in turn and find a ball in C_2 containing $f(x_i)$ and let $c_2(x_i)$ denote the ball's center. Write down $c_2(x_i)$ as an index into C_2. This costs $\Delta m\lg(1+4/\varepsilon)$ bits when summed over all x_i. Next, let A

denote the $\Delta \times m$ matrix having one row per x_i, where the i'th row is $c_2(x_i)$. Let W denote the subspace of $\mathbb{R}^\Delta$ spanned by the columns of A. We have $\dim(W) \leq m$. Define T as the convex body

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

That is, T is the intersection of the subspace W with the Δ-dimensional ℓ_∞ unit ball B_∞^Δ. Now let C_∞ be a minimum cardinality covering of $(2^{12}\varepsilon)T$ 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 $|C_\infty| \leq 2^{m \lg(1+2^{13})}$. To encode the vectors $y_1, \ldots, y_Q$ we make use of the following lemma, whose proof we give in Section 5.1:

Lemma 5. *For every x_j and y_i in P, we have*

$$|\langle c_2(x_j), f(y_i)\rangle - \langle x_j, y_i\rangle| \leq 10\varepsilon.$$

From Lemma 5 and Corollary 2, it follows that $|\langle c_2(x_j), f(y_i)\rangle| \leq 10\varepsilon + 2^{11}\varepsilon < 2^{12}\varepsilon$ for every x_j and y_i in P. Since the j'th coordinate of $Af(y_i)$ equals $\langle c_2(x_j), f(y_i)\rangle$, it follows that $Af(y_i) \in (2^{12}\varepsilon)T$. Using this fact, we encode each y_i by finding some vector $c_\infty(y_i)$ such that $c_\infty(y_i) + \varepsilon T$ is a convex shape in the covering C_∞ and $Af(y_i) \in c_\infty(y_i) + \varepsilon T$. We write down $c_\infty(y_i)$ as an index into C_∞. This costs a total of $Qm \lg(1 + 2^{13}) < 14Qm$ bits over all y_i. We now describe our decoding algorithm.

Decoding Algorithm. To recover $P = \{0, x_1, \ldots, x_\Delta, y_1, \ldots, y_Q\}$ from the above encoding, we only have to recover $y_1, \ldots, y_Q$ as $\{0, x_1, \ldots, x_\Delta\}$ 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 $c_2(x_i)$ for $i = 1, \ldots, \Delta$. 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 $c_\infty(y_1), \ldots, c_\infty(y_Q)$. What remains is to recover $y_1, \ldots, y_Q$. Since y_i is uniquely determined from the set $S_i \subseteq \{1, \ldots, \Delta\}$ of k indices, we focus on recovering this set of indices for each y_i.

For $i = 1, \ldots, Q$ recall that $Af(y_i)$ is in $c_\infty(y_i) + \varepsilon T$. Observe now that:

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

But the j'th coordinate of $Af(y_i)$ is $\langle c_2(x_j), f(y_i)\rangle$. We combine the above with Lemma 5 to deduce $|(c_\infty(y_i))_j - \langle x_j, y_i\rangle| \leq 11\varepsilon$ for all j. From Corollary 2, it follows $(c_\infty(y_i))_j < 12\varepsilon$ for $j \notin S_i$ and $(c_\infty(y_i))_j > 2^7\varepsilon$ for $j \in S_i$. We finally conclude that the set S_i, and thus y_i, is uniquely determined from $c_\infty(y_i)$.

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 $\Delta m \lg(1 + 4/\varepsilon) + 14Qm \leq \Delta m \lg(5/\varepsilon) + 14Qm < \Delta m \lg(1/\varepsilon) + 17nm$ bits. We chose $\Delta = n/\lg(1/\varepsilon)$ and the number of bits is thus no more than $18nm$. But $|\mathcal{P}| \geq \left(\binom{\Delta}{k}/2\right)^Q \geq (\Delta/(2k))^{kQ} = (\Delta/(2k))^{k(n-\Delta-1)} \geq (\Delta/(2k))^{kn/2}$. We therefore must have

$$\begin{aligned} 18nm &\geq (kn/2)\lg(\Delta/(2k)) \Rightarrow \\ m &= \Omega(\varepsilon^{-2}\lg(\varepsilon^2 n/\lg(1/\varepsilon))). \end{aligned}$$

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

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

This concludes the proof of Theorem 2.

5.1 Proof of Lemma 5

In this section, we prove the lemma:

Restatement of Lemma 5. *For every x_j and y_i in P, we have*

$$|\langle c_2(x_j), f(y_i)\rangle - \langle x_j, y_i\rangle| \leq 10\varepsilon.$$

Proof. First note that:

$$\begin{aligned}\langle c_2(x_j), f(y_i)\rangle &= \langle c_2(x_j) - f(x_j) + f(x_j), f(y_i)\rangle \\ &= \langle f(x_j), f(y_i)\rangle + \langle c_2(x_j) - f(x_j), f(y_i)\rangle \\ &\in \langle f(x_j), f(y_i)\rangle \pm \|c_2(x_j) - f(x_j)\|_2\|f(y_i)\|_2.\end{aligned}$$

Since C_2 was a covering with εB_2^m, we have $\|c_2(x_j) - f(x_j)\|_2 \leq \varepsilon$. Furthermore, since f satisfies (3), we have from Corollary 2 and $0 \in P$ that $\|f(y_i)\|_2^2 \leq 2(1+\varepsilon) \Rightarrow \|f(y_i)\|_2^2 \leq 4 \Rightarrow \|f(y_i)\|_2 \leq 2$. We thus have:

$$\langle c_2(x_j), f(y_i)\rangle \in \langle f(x_j), f(y_i)\rangle \pm 2\varepsilon. \tag{4}$$

To bound $\langle f(x_j), f(y_i)\rangle$, observe that

$$\|f(x_j) - f(y_i)\|_2^2 = \|f(x_j)\|_2^2 + \|f(y_i)\|_2^2 - 2\langle f(x_j), f(y_i)\rangle.$$

This implies that

$$\begin{aligned}2\langle f(x_j), f(y_i)\rangle &\in \|x_j\|_2^2(1\pm\varepsilon) + \|y_i\|_2^2(1\pm\varepsilon) - \|x_i - y_i\|_2^2(1\pm\varepsilon) \\ &\subseteq 2\langle x_j, y_i\rangle \pm \varepsilon(\|x_j\|_2^2 + \|y_i\|_2^2 + \|x_j - y_i\|_2^2) \\ &\subseteq 2\langle x_j, y_i\rangle \pm \varepsilon(4(\|x_j\|_2^2 + \|y_i\|_2^2))\end{aligned}$$

That is, we have

$$\langle f(x_j), f(y_i)\rangle \in \langle x_j, y_i\rangle \pm 2\varepsilon(\|x_j\|_2^2 + \|y_i\|_2^2).$$

Using Corollary 2, we have

$$\begin{aligned}\langle f(x_j), f(y_i)\rangle &\in \langle x_j, y_i\rangle \pm 2\varepsilon((1+\varepsilon/2)+2) \\ &\subseteq \langle x_j, y_i\rangle \pm 8\varepsilon.\end{aligned}$$

Inserting this in (4), we obtain

$$\langle c_2(x_j), f(y_i)\rangle \in \langle x_j, y_i\rangle \pm 10\varepsilon.$$

□

6 Proofs of Lemmas on Nearly Orthogonal Vectors

In the following, we prove the lemmas from Section 4.

Restatement of Lemma 3. *For any $0 < \mu < 1$, $1 \leq N \leq \max\{d, e^{O(\mu^2 d)}\}$ and integer $1 \leq k \leq N$, there exists $(X, \mathcal{F})$ with $X \subset \mathbb{R}^d$ and $|X| = N$, such that $(X, \mathcal{F})$ is k-wise μ-incoherent and*

$$|\mathcal{F}| \geq \binom{N}{k}/2.$$

Proof. When d is the maximum in the bound $N \leq \max\{d, e^{O(\mu^2 d)}\}$, the lemma follows by setting $X = \{e_1, \dots, e_N\}$ and $\mathcal{F}$ the collection of all k-sized subsets of X.

For the other case, we prove the lemma by letting X be a set of N independent gaussian vectors in $\mathbb{R}^d$ each with identity covariance matrix, scaled by a factor $1/\sqrt{d}$, i.e. each coordinate of each vector is independently $\mathcal{N}(0, 1/d)$ distributed. We then show that as long as μ, N and k satisfy the requirements of the lemma, then with non-zero probability, we can find $\mathcal{F}$ of at least $\binom{N}{k}/2$ k-sized subsets of X which makes $(X, \mathcal{F})$ k-wise μ-incoherent.

Let $x_1, x_2, \dots, x_N$ denote the vectors in X and consider any fixed x_i. First note that $\|x_i\|_2^2 \sim (1/d)\chi_d^2$. We thus get from tail bounds on the chi-squared distribution that

$$\mathbb{P}(|\|x_i\|_2^2 - 1| > \mu) < e^{-\Omega(\mu^2 d)}.$$

Let E_i denote the event $|\|x_i\|_2^2 - 1| \leq \mu$. Consider now any subset $S \subseteq [N]$ with $|S| \leq k$. Let $x_S = \sum_{j \in S} x_j/\sqrt{k}$. For any fixed vector $y \in \mathbb{R}^d$, we have

$$\langle y, x_S \rangle = \sum_{i=1}^{d} \sum_{j \in S} y_i (x_j)_i/\sqrt{k} = \sum_{i=1}^{d} y_i \sum_{j \in S} (x_j)_i/\sqrt{k}.$$

For every i, $\sum_{j \in S}(x_j)_i/\sqrt{k}$ is $\mathcal{N}(0, |S|/(dk))$ distributed and these are independent across the different i's. Thus $\langle y, x_S \rangle \sim \mathcal{N}(0, \|y\|_2^2 |S|/(dk))$. Therefore, if y is a vector with $\|y\|_2^2 \leq 1 + \mu$, then

$$\mathbb{P}(|\langle y, x_S \rangle| > \mu) < e^{-\Omega(\mu^2 dk/(|S|\|y\|_2^2))} = e^{-\Omega(\mu^2 d)}.$$

Now fix an x_i and a set $S \subseteq [N]$ with $|S| = k$. Observe that if we condition on the event E_i, all vectors x_h with $h \in S \setminus \{i\}$ are still independent standard gaussians scaled by $1/\sqrt{d}$. We thus have:

$$\mathbb{P}(|\langle x_i, \sum_{j \in S : i \neq j} x_j/\sqrt{k} \rangle| > \mu \mid E_i) < e^{-\Omega(\mu^2 d)}.$$

We say that S fails if there is some $x_i \in X$ such that either

1. $|\|x_i\|_2^2 - 1| > \mu$.
2. $|\langle x_i, \sum_{j \in S : i \neq j} x_j/\sqrt{k} \rangle| > \mu$.

By a union bound, the first event happens with probability at most $Ne^{-\Omega(\mu^2 d)}$. For the second event, consider a fixed x_i. Then

$$\begin{aligned}\mathbb{P}(|\langle x_i, \sum_{j \in S : i \neq j} x_j/\sqrt{k} \rangle| > \mu) &\leq \mathbb{P}(\neg E_i) + \mathbb{P}(|\langle x_i, \sum_{j \in S : i \neq j} x_j/\sqrt{k} \rangle| > \mu \mid E_i) \\ &\leq e^{-\Omega(\mu^2 d)} + e^{-\Omega(\mu^2 d)}.\end{aligned}$$

Thus we can again union bound over all x_i, and conclude that S fails with probability at most $Ne^{-\Omega(\mu^2 d)}$. As long as $N \leq e^{\gamma \mu^2 d}$ for a sufficiently small constant $\gamma > 0$, this probability is less than $1/2$. Therefore the expected number of sets S that do not fail is at least $\binom{N}{k}/2$. This means that there must exist a choice of X for which there are at least $\binom{N}{k}/2$ sets S that do not fail. This shows the existence of the claimed $(X, \mathcal{F})$ since as long as at least one S does not fail, $(X, \mathcal{F})$ also satisfies $|\|x_i\|_2^2 - 1| \leq \mu$ for all $x_i \in X$. $\square$

Restatement of Lemma 4. *Let $X = \{x_1, \dots, x_N\}$ and let $(X, \mathcal{F})$ be k-wise μ-incoherent for some $0 < \mu < 1$ and $k \geq 1$. Let $S \in \mathcal{F}$ and define $y = \sum_{j \in S} x_j/\sqrt{k}$. Then y satisfies:*

1. $\|y\|_2^2 \leq 1 + (\sqrt{k} + 1)\mu$.

2. *For a vector $x_j \in X$ such that $j \notin S$, we have $|\langle y, x_j\rangle| \le \mu$.*

3. *For a vector $x_j \in X$ such that $j \in S$, we have $(1-\mu)/\sqrt{k} - \mu \le |\langle y, x_j\rangle| \le (1+\mu)/\sqrt{k} + \mu$.*

Proof. For the first item, observe that

$$\begin{aligned}\|y\|_2^2 &= \sum_{i\in S}\sum_{j\in S}\langle x_i, x_j\rangle/k \\ &= \sum_{i\in S}\|x_i\|_2^2/k + \sum_{i\in S}\sum_{j\in S: i\neq j}\langle x_i, x_j\rangle/k \\ &\le 1+\mu+1/\sqrt{k}\sum_{i\in S}\left|\langle x_i, \sum_{j\in S:i\neq j} x_j/\sqrt{k}\rangle\right| \\ &\le 1+\mu+\sqrt{k}\mu.\end{aligned}$$

For the second item, let $x_j \in X$ with $j \notin S$. Then

$$\begin{aligned}|\langle y, x_j\rangle| &= \left|\langle x_j, \sum_{i\in S} x_i/\sqrt{k}\rangle\right| \\ &\le \mu.\end{aligned}$$

For the third term, let $j \in S$. Then

$$\begin{aligned}|\langle y, x_j\rangle| &= \left|\langle x_j, \sum_{i\in S} x_i/\sqrt{k}\rangle\right| \\ &= \left|\|x_j\|_2^2/\sqrt{k} + \langle x_j, \sum_{i\in S:i\neq j} x_i/\sqrt{k}\rangle\right|.\end{aligned}$$

This shows that

$$|\langle y, x_j\rangle| \ge (1-\mu)/\sqrt{k} - \mu.$$

and also that

$$|\langle y, x_j\rangle| \le (1+\mu)/\sqrt{k} + \mu.$$

□

References

[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.

[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.

[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.