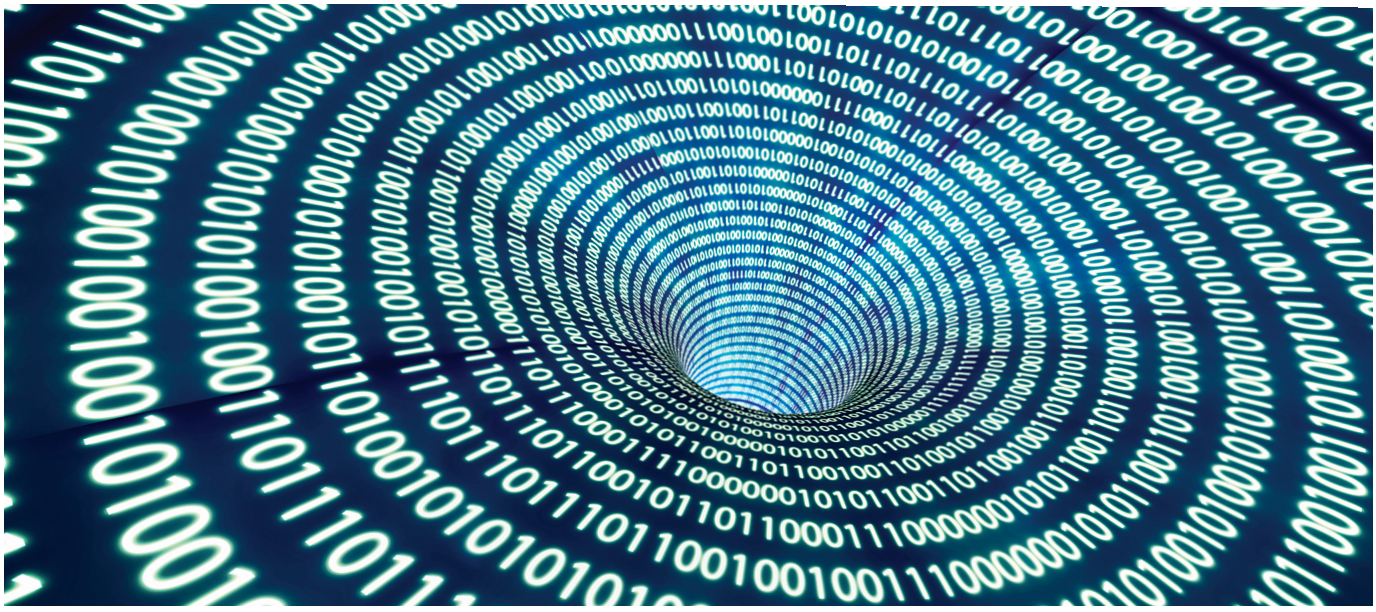
Dimensionality Reduction in Euclidean Space



Jelani Nelson

I begin with a description of what this article is *not* about. It is not about Principal Component Analysis (PCA), Kernel PCA, Multidimensional Scaling, ISOMAP, Hessian Eigenmaps, or other methods of dimensionality reduction primarily created to help understand high-dimensional datasets. Rather, this article focuses on high dimensionality as a barrier to algorithmic efficiency (i.e., low running time and/or memory consumption), and explores how dimension reduction can be used as an *algorithmic tool* to overcome this barrier. In fact, as we discuss more in length in Section 5.2, this view is not only different than but complementary to the above-mentioned approaches, as the form of dimension reduction we focus on here for example can be used to obtain faster algorithms for approximate PCA.

Moving back a few steps from dimension reduction, more generally an effective technique in the design of algorithms processing geometric data is to employ a *metric embedding* to transform the input in one given metric space to another that is computationally friendlier, and then to work over the latter space (see the survey [Ind01]). To measure the quality of such an embedding, we use the following terminology: given a *host metric space* $\mathcal{X} = (X, d_X)$ and a *target space* $\mathcal{Y} = (Y, d_Y)$, $f : X \to Y$ is said to be a *bi-Lipschitz embedding with distortion D* if there exists a (scaling) constant c such that for all $x, y \in X$,

$$c \cdot d_X(x, y) \le d_Y(f(x), f(y)) \le cD \cdot d_X(x, y). \qquad (1)$$

To illustrate the embedding paradigm in action, consider the *k-median* problem. The input is a finite metric space $\mathcal{X} = (X, d_X)$, $|X| = n$, together with an integer $1 \le k \le n$. The goal is to compute

$$S^* = \operatorname*{argmin}_{\substack{S \subset X \\ |S|=k}} \sum_{x \in X} \min_{c \in S} d_X(x, c). \qquad (2)$$

Jelani Nelson is a professor of electrical engineering and computer science at the University of California, Berkeley. His email address is minilek@berkeley.edu.

The author's research was supported by NSF award CCF-1951384, ONR grant N00014-18-1-2562, ONR DORECG award N00014-17-1-2127, an Alfred P. Sloan Research Fellowship, and a Google Faculty Research Award.

Due to publisher constraints, only a limited number of references could be included. For a version of this article with a full list of references, please see full version on the arXiv.

Communicated by Notices *Associate Editor Reza Malek-Madani.*

DOI: https://doi.org/10.1090/noti2166

That is, we would like to partition X into k clusters, together with identifying a cluster center c in each cluster, so as to minimize the sum of distances from every $x \in X$ to its closest cluster center. If $\mathcal{X}$ can be an arbitrary n-point metric space, then this problem is known to be NP-hard. Meanwhile when $\mathcal{X}$ is the shortest path metric on a tree, the problem can be solved exactly in time $O(kn^2)$ via the Kariv-Hakimi dynamic programming algorithm.[1] Tree shortest path metrics are thus an example of what we would call a computationally friendly metric space for the k-median problem. Thus if $\mathcal{X}$ admits an algorithmically efficient embedding into a tree metric with some small distortion D, we can obtain a fast D-approximation algorithm for k-median on $\mathcal{X}$ (i.e., achieving a clustering cost that is at most a factor D larger than optimal) by first embedding our original metric into some tree T and then solving k-median exactly in T. In fact it has been shown by Fakcharoenphol et al., following previous work of Bartal, that any n-point metric space embeds into a distribution over tree metrics with distortion $O(\log n)$. We will not discuss here what distortion means for probabilistic embeddings into a *distribution* over target spaces, but to make our case for the embedding paradigm it suffices to point out that these results implied the first ever polynomial time algorithms for k-median computation in arbitrary metric spaces with approximation factor at most polylogarithmic in n.

In this article we focus on embeddings in which both the host and target spaces are normed spaces, in which case we can drop the scaling factor c in equation (1). We even more specifically focus on the case when $\mathcal{X}, \mathcal{Y}$ are finite-dimensional subspaces of the same normed space $\mathcal{Z}$, and where $dim(\mathcal{Y}) \leq dim(\mathcal{X})$ so that f provides us with the algorithmic advantage of *dimension reduction*. As one might imagine, several algorithms for high-dimensional computational geometry problems have running times or memory requirements which grow (sometimes poorly) with the dimension of the input. An example is the nearest neighbor search data structural problem, in which one wants to preprocess a set of input points $x_1, \ldots, x_n \in \mathbb{R}^d$ to create a low-memory data structure $\mathcal{D}$ such that later one can quickly identify the closest x_i to some query point $q \in \mathbb{R}^d$ by querying $\mathcal{D}$.[2] The best known algorithms for this problem with fast query time (in terms of n) either have running time or memory usage exponential in d (see the discussion in [HPIM12]).

A natural question is then: for which normed spaces do there exist such dimensionality-reducing maps with low distortion? An early and seminal result in this direction was given by Johnson and Lindenstrauss [JL84], who showed that near-isometric embeddings exist when $\mathcal{X}, \mathcal{Y}$ are Euclidean.

Lemma 1 (JL lemma [JL84]). *Let $\varepsilon \in (0,1)$ and $X \subset \mathbb{R}^d$ be arbitrary with $|X|$ having size $n > 1$. Then there exists $f : X \to \mathbb{R}^m$ with $m = O(\varepsilon^{-2} \log n)$ such that for all $x, y \in X$,*

$$\|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \varepsilon)\|x - y\|_2. \quad (3)$$

In fact, all proofs of the JL lemma show that f can be taken as a linear map. The various known proofs of the JL lemma all identify a distribution Γ over $\mathbb{R}^{m \times d}$ such that if one draws a random $\Pi \sim \Gamma$, then $f(x) = \Pi x$ satisfies equation (3) with high probability. In the original proof [JL84], Γ was taken as a scaled orthogonal projection onto a random m-dimensional subspace of $\mathbb{R}^d$ (and hence their technique is often called the *random projection method*), though since then several other distributions have been shown to provide a similar guarantee.

Hearing of such a result naturally inspires certain follow-up questions. Is low-distortion dimension reduction possible in other normed spaces, e.g., ℓ_p for $p \neq 2$? Is the $m = O(\varepsilon^{-2} \log n)$ bound in the JL lemma the best possible? Is it possible to obtain a distribution Γ providing the JL lemma as mentioned above such that $\Pi \sim \Gamma$ can be sampled using few random bits? Given that the stated primary motivation of dimension reduction is algorithmic efficiency, just how fast can the mapping $x \mapsto \Pi x$ be performed?

1. Dimension Reduction in Other Spaces

Given the dimension reduction possible in Euclidean space, one might wonder in what other spaces such a result is possible. A negative result was proven by Johnson and Naor, who showed that at least for *linear* embeddings, spaces that enjoy dimension reduction as good as in the Euclidean case must themselves be nearly Euclidean.

Theorem 1 ([JN10]). *Suppose Z is normed space satisfying the property that for every $X \subset Z$, $|X| = n$, there exists a linear mapping $f : Z \to E$ for an $O(\log n)$-dimensional subspace $E \subset Z$ such that f has $O(1)$-distortion when restricted to X. Then, every k-dimensional linear subspace of Z embeds into Euclidean space with distortion $2^{2^{O(\log^* k)}}$.*

In the above, $\log^* m$ is the number of times one must take the iterated logarithm of m, base two, to obtain a number which is at most 1. For example, $\log^*(2^{2^{2^2}}) = 4$. The key takeaway here is that $\log^* m$ is a very slow-growing function, so that the distance to being Euclidean is small.

The theorem though does not preclude the existence of some form of dimension reduction in spaces that are not

[1] *We use standard asymptotic notation. For functions f, g: $f = O(g)$ if $\limsup_{x \to \infty} |f(x)/g(x)| < \infty$. $f = \Omega(g)$ if $g = O(f)$; $f = \Theta(g)$ if both $f = O(g)$ and $f = \Omega(g)$; $f = o(g)$ if $\lim_{x \to \infty} f(x)/g(x) = 0$; and $f = \omega(g)$ if $g = o(f)$.*

[2] *Though specifically for the nearest neighbor problem, an embedding satisfying a weaker guarantee suffices for applications.*

nearly Euclidean. In particular, one can still shoot for dimension reduction bounds that are $\omega(\log n)$, or potentially achieve $O(\log n)$ target dimension with $O(1)$ distortion via nonlinear embeddings. Several results exist showing that some nontrivial dimension reduction in ℓ_p-spaces, for example, is possible. On the negative side, Brinkman and Charikar have shown that for an n-point set endowed with the ℓ_1 metric, which is known to always be embeddable isometrically into dimension $\binom{n}{2}$, any embedding into ℓ_1 with constant distortion D must have embedding dimension $m = n^{\Omega(1/D^2)}$. For near isometries ($D = 1+\varepsilon$), Andoni et al. showed $m = n^{1-O(1/\log(1/\varepsilon))}$ is required. Meanwhile, the best known upper bound is $m = O(n/\varepsilon^2)$ by Newman, building upon the sparsification technique of Batson, Spielman, and Srivastava. For ℓ_p for even integer p, Schechtman obtained the bound $m = C_p n^{p/2}/\varepsilon^2$ with $C_p = O(p^{-p/2})$. For the Schatten-1 norm, also known as the nuclear norm, Naor, Pisier, and Schechtman showed that constant distortion into $m = n^{o(1)}$ is impossible.

2. Sharpness of the Johnson-Lindenstrauss Lemma

The original paper of Johnson and Lindenstrauss [JL84] proving the JL lemma showed a lower bound on the optimal target dimension m to achieve equation (3) via a volume argument. The argument is succinct enough that we will repeat it here. Consider the set X of $n+1$ points $0, e_1, \dots, e_n \in \mathbb{R}^n$, where e_i is the ith standard basis vector. Let f be a $(1+\varepsilon)$-distortion embedding of X into $\mathbb{R}^m$ for $\varepsilon < 1/2$, where we may assume $f(0) = 0$ by translation. Then since f preserves distances to 0, we must have $\|f(e_i)\|_2 \le (1+\varepsilon)\|e_i\|_2 < 3/2$ for all i, so that a radius-1/2 ball about $f(e_i)$ lies entirely within the radius-2 ball about the origin in $\mathbb{R}^m$. We must also have $\|f(e_i) - f(e_j)\|_2 > (1-\varepsilon)\sqrt{2} > 1/2$, so that the radius-1/4 balls B_i about the $f(e_i)$ are disjoint. Thus we have n radius-1/4 balls $B_1, \dots, B_n$ that all lie entirely within a radius-2 ball but are disjoint. Letting $B_{\ell_2^m}(r)$ denote the radius-r ball about the origin in $\mathbb{R}^m$,

$$\begin{aligned} vol(B_{\ell_2^m}(2)) &\ge vol\Big(\bigcup_i B_i\Big) \\ &= \sum_{i=1}^n vol(B_i) \\ &= n \cdot vol(B_{\ell_2^m}(1/4)). \end{aligned}$$

Thus

$$n \le \frac{vol(B_{\ell_2^m}(2))}{vol(B_{\ell_2^m}(1/4))} = 8^m,$$

so that $m \ge \log_8 n = \Omega(\log n)$.

Unfortunately, the above approach does not extend to show that m must grow by more than a constant factor beyond $\log_8 n$ as $\varepsilon \to 0$. Subsequently, Alon showed the lower bound $\Omega(\varepsilon^{-2}\log n/\log(1/\varepsilon))$ for $\varepsilon > 1/\sqrt{n}$. Roughly, the approach was to let X be as above (0, together with the simplex), and to again let f be a low-distortion embedding as above with $f(0) = 0$. Then Alon defined a matrix $B \in \mathbb{R}^{m\times n}$ whose ith column is $b_i = f(e_i)/\|f(e_i)\|_2$. Thus, the column norms of B are 1, and one can show that $\|b_i - b_j\|_2 = (1+O(\varepsilon))\sqrt{2}$ implies that the pairwise dot products between the b_i are each $O(\varepsilon)$. Therefore, $A = B^\top B$ is a "near-identity" matrix: its diagonal entries are all 1, and off the diagonal all entries are $O(\varepsilon)$. Alon showed that if such a matrix has off-diagonal entries at most $1/\sqrt{n}$, then $rank(A) = \Omega(n)$, implying $m = \Omega(n)$ since $m \ge rank(B) = rank(A)$. Of course though our A does not necessarily have $\varepsilon < 1/\sqrt{n}$; ε is whatever it is! But if one defines $A^{\otimes r}$ to be the matrix with $(A^{\otimes r})_{i,j} = (A_{i,j})^r$, then $A^{\otimes r}$ *does* have this property for $r = \lceil \log(\sqrt{n})/\log(1/\varepsilon)\rceil$. One then applies the rank lower bound to this matrix to say $rank(A^{\otimes r}) = \Omega(n)$, combined with an inequality upper bounding $rank(A^{\otimes r})$ in terms of m, r.

Progress halted after Alon's lower bound for some time, and in particular there was not even a known candidate for a set X for which the JL bound was sharp (for the simplex, it was known even to Alon that better m was achievable when $\varepsilon < \exp(-c\sqrt{\log n})$). Some progress came eventually via a result of Jayram and Woodruff, with a later alternate proof by Kane, Meka, and Nelson, that the *distributional* JL lemma is optimal.

Definition 1. A distribution Γ over $\mathbb{R}^{m\times d}$ is an (ε,δ)-JL distribution if

$$\forall x \in \mathbb{R}^d,\ \mathop{\mathbb{P}}_{\Pi\sim\Gamma}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta.$$

Lemma 2 (Distributional JL lemma [JL84]). *For all $\varepsilon,\delta \in (0,1)$ and integer $d > 1$, there exists an (ε,δ)-JL distribution with $m = O(\varepsilon^{-2}\log(1/\delta))$.*

All proofs of the JL lemma are via the distributional JL lemma, taking $\delta < 1/\binom{n}{2}$ and then union bounding to argue that for $\Pi \sim \Gamma$, $\|\Pi z\|_2 \approx \|z\|_2$ for all $z \in X - X$ simultaneously, and it has been shown that any (ε,δ)-distribution must have $m = \Omega(\min\{d, \varepsilon^{-2}\log(1/\delta)\})$, which is sharp since for $\varepsilon^{-2}\log(1/\delta) > d$ one can instead take Γ supported only on the identity map. The proof of Jayram and Woodruff was via a communication complexity argument, whereas the proof of Kane et al. was via Yao's minimax principle. Specifically for the latter, if $\mathcal{D}$ is an arbitrary distribution over points in $\mathbb{R}^d$ and Γ is an (ε,δ)-JL distribution,

we have

$$\begin{aligned}
&\forall x \in \mathbb{R}^d, \ \mathop{\mathbb{P}}_{\Pi\sim\Gamma}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta \\
\implies &\mathop{\mathbb{P}}_{x\sim\mathcal{D}} \mathop{\mathbb{P}}_{\Pi\sim\Gamma}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta \\
\implies &\mathop{\mathbb{P}}_{\Pi\sim\Gamma} \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta \\
\implies &\exists \Pi \in \mathbb{R}^{m\times d} \mathop{\mathbb{P}}_{x\sim\mathcal{D}}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta.
\end{aligned}$$

One can then show the final statement is impossible unless $m = \Omega(\min\{d, \varepsilon^{-2}\log(1/\delta)\})$ for $\mathcal{D}$ being the uniform distribution on the sphere.

Of course a sharp lower bound on JL distributions does not imply a sharp lower bound for Euclidean dimensionality reduction, since in principle there could be a way of constructing optimal embeddings that does not use JL distributions at all (and in fact could use nonlinear embeddings!). While the later work of Larsen and Nelson did not rule out the latter possibility of better nonlinear embeddings, it did show that the JL lemma is sharp for Euclidean dimension reduction if one is only allowed to use linear embeddings. The basic idea was simple: let X be the union of $\{0, e_1, \ldots, e_d\}$ and $n - d - 1$ independent gaussian vectors g_j, and let $\Pi \in \mathbb{R}^{m\times d}$ be arbitrary. Then we need $\|\Pi e_i\|_2 \approx 1$ for all i, so the Frobenius norm of Π should be $O(\sqrt{d})$. We also must have $\|\Pi g_j\|_2 \approx \|g_j\|_2$ for all j. But for fixed j, this fails to hold with some probability just due to random fluctuation, and thus if n is large enough this will fail to hold for some j with high probability. If we then union bound over *all* Π in some fine enough finite covering of the set of all $\Pi \in \mathbb{R}^{m\times d}$ with $O(\sqrt{d})$ bounded Frobenius norm, we can argue that with positive probability X is a hard set for all such Π simultaneously, and a standard approach can then pass the hardness of X onto all Π of bounded Frobenius norm and not just those in the covering.

Finally, in a later work of Larsen and Nelson [LN17], the optimality of the JL lemma was shown even amongst nonlinear embeddings. Specifically, it was shown that for any $\varepsilon > 1/\min\{n, d\}^{0.499}$, there exists a point set $X \subset \mathbb{R}^d$, $|X| = n$, such that any $(1+\varepsilon)$-distortion embedding into ℓ_2^m requires $m = \Omega(\varepsilon^{-2}\log n)$. This lower bound restriction on ε is close to necessary, since it amounts to requiring $1/\varepsilon^2 < \min\{n, d\}^{0.998}$. Note that one must require at least $1/\varepsilon^2 < \min\{n, d\}$ for JL to be optimal, since there is an isometric embedding of X into dimension $\min\{d, n-1\}$. One can embed into dimension d using the identity map, and into distortion $n - 1$ by noticing that X spans an at most $(n-1)$-dimensional subspace (once we translate one of the points to the origin, which does not affect distances). A subsequent work of Alon and Klartag gave sharper bounds for ε approaching this boundary, and in particular gave a lower bound of $m = \Omega(\min\{n, d, \varepsilon^{-2}\log(\varepsilon^2 n)\})$.

The method of proof in [LN17] was via a counting argument. A collection $\mathcal{X}$ of point sets in $\mathbb{R}^d$, each of size n, is defined with the following property: if for every $X \in \mathcal{X}$ there is a $(1+\varepsilon)$-distortion embedding of X into $\mathbb{R}^m$, then there is an injection from $\mathcal{X}$ to $\{0,1\}^{Cnm}$ for some universal constant C. Thus, we obtain the lower bound that some $X \in \mathcal{X}$ must require $m \geq C^{-1}(\log|\mathcal{X}|)/n$. The construction of this injection uses that if f preserves the norms of $x, y, x - y$ up to $1 - \varepsilon$ and x, y are in the unit Euclidean ball, then $\langle x, y\rangle$ and $\langle f(x), f(y)\rangle$ must differ by only an additive $O(\varepsilon)$. The injection is defined via an encoding based on the (rounded) dot products of embeddings of certain pairs of vectors in X, and ultimately the encoding scheme can be viewed as providing a lower bound on the so-called *packing number* of the set of all rank-m $n \times n$ Gram matrices obtained from sets of unit vectors. The packing number $\mathcal{M}(T, d, \varepsilon)$ is the maximum number of disjoint radius-ε balls under metric d that can be obtained with centers in T, which is related to the *covering* number $\mathcal{N}(T, d, \varepsilon)$, which is the minimum number of radius-ε balls in metric d centered at points in T required such that every $t \in T$ is contained in (or "covered" by) at least one ball. This connection between the argument of [LN17] and lower bounding such packing numbers was made explicit later by Alon and Klartag, in which an alternate covering upper bound proof was also given (and which could recover tighter bounds as $\varepsilon \to 1/\sqrt{n}$). The work of Alon and Klartag also gives a memory-efficient data structure for querying approximate dot products of pairs of vectors in an input database of vectors (see also the work of Indyk and Wagner, for approximate distance query data structures using low memory).

3. Randomness-Efficiency: Sampling the Embedding

Just as time and memory are computational resources whose consumption a computer scientist aims to minimize in the development of algorithms, randomness is also a similar such resource, as preparing a source of random bits to feed into an algorithm requires effort. Indeed, the computational field of *pseudorandomness* focuses exactly on this resource, by studying the question: how can we design algorithms to stretch s truly random bits into $N \gg s$ pseudorandom bits that look "random enough" to some class of algorithms that they perform nearly just as well as if the N bits had been uniform independent random bits? See for example the book on this subject by Vadhan [Vad11].

Coming back to the JL lemma, recall the distributional JL lemma, Lemma 2. A natural question then is: what is the fewest number of random bits b such that one can design an (ε, δ)-JL distribution Γ and an algorithm $\mathcal{A} : \{0,1\}^b \to \mathbb{R}^{m\times d}$ such that $\mathcal{A}(\mathcal{U}_b)$ is distributed as Γ? Here $\mathcal{U}_b$ is the uniform distribution on $\{0,1\}^b$.

reference	# random bits	description of Π
[Ach03]	md	i.i.d. Rademacher entries
[CW09]	$O(\log(1/\delta)\log d)$	k-wise independent Rademacher entries
[KMN11]	$O(\log d + \log(1/\delta)(\log\log(1/\delta) + \log(1/\varepsilon)))$	geometrically increasing independence

Figure 1. Methods to sample from JL distributions using few random bits.

Before delving into the state-of-the-art results addressing this question, we take a digression to introduce k-wise independent sample spaces.

Definition 2. For a finite set S, consider a distribution $\mathcal{D}$ generating $(s_1, \dots, s_n) \in S^n$. Let $\mathcal{E}_{i,t}$ be the event $s_i = t$. Then, we say $\mathcal{D}$ is a *k-wise independent sample space* if for any $1 \le i_1 < i_2 < \cdots < i_k \le n$ and any $t_1, \dots, t_k \in S$,

$$\mathop{\mathbb{P}}_{s\sim\mathcal{D}}\left(\bigwedge_{j=1}^{k} \mathcal{E}_{i_j,t_j}\right) = \frac{1}{|S|^k}.$$

In other words, $\mathcal{D}$ is a k-wise independent sample space if when we look at any k coordinates of s at a time, the marginal distribution of just those k coordinates is uniform in S^k.

In the language of computer science, one often devises such sample spaces by constructing a family $\mathcal{H}$ of functions mapping $[n]$ to S (a so-called "hash family"), selecting $h \in \mathcal{H}$ uniformly at random, and then setting $s_i := h(i)$. The function h is typically called a *hash function*. The case that $\mathcal{D}$ samples uniformly in S^n corresponds to $\mathcal{H}$ being the set of all $|S|^n$ functions mapping $[n]$ to S, which is certainly k-wise independent for any k. When $S = [n]$ and $n = 2^\ell$, it was shown by Carter and Wegman that the following hash family is also k-wise independent:

$$\mathcal{H}_{\text{poly}} = \left\{ h(x) = \sum_{i=0}^{k-1} a_i x^i \,:\, a_0, \dots, a_{k-1} \in \mathbb{F}_n \right\},$$

where all arithmetic in the definition of $h(x)$ is over the field $\mathbb{F}_n$. Note a sample from this sample space can be generated using only $O(k \log n)$ random bits, as opposed to $O(n \log n)$ needed for a uniformly random element of S^n. Note also that if $S = \{0, 1\}$, we can use the same construction but where now we set $s_i = h(i)|_0$, i.e., put $\mathbb{F}_n$ in bijective correspondence with binary strings of length $\log_2 n$ and then project $h(i)$ to its least significant bit (its "0th bit") in binary.

The key reason we introduced k-wise independent sample spaces is that if $s = (s_1, \dots, s_n)$ is drawn from such a distribution, then for any degree at most k polynomial p, $\mathbb{E}_s\, p(s)$ is equal to the case of s being a uniform sample in S^n (seen by expanding $p(x)$ into a sum of monomials and observing that each monomial's expectation is preserved by k-wise independence). Next, the typical way a distribution is shown to satisfy the distributional JL lemma is by bounding the moments of $Z := |\|\Pi x\|_2^2 - \|x\|_2^2|$, either directly or indirectly, via bounding its moment generating function $\mathbb{E}\, e^{tZ}$ and then applying Markov's inequality. Now observe that for k an even integer, Z^k is a degree-$2k$ polynomial in the entries of Π, and thus if Π has i.i.d. entries, this kth moment is determined by $2k$-wise independence.

The results. Achlioptas showed that Π having i.i.d. entries in $\{-1/\sqrt{m}, 1/\sqrt{m}\}$ for $m = O(\varepsilon^{-2}\log(1/\delta))$ provides an (ε, δ)-JL distribution, via bounding the moment-generating function (and hence moments), and in fact the JL distribution property holds by analyzing the kth moment for $k = O(\log(1/\delta))$ (shown by Clarkson and Woodruff). Thus, Π from a JL distribution can be sampled using only $O(\log(1/\delta)\log(md)) = O(\log(1/\delta)\log d)$ bits. Alternatively, one can use i.i.d. entries and sample Π using $md = O(\varepsilon^{-2} d \log(1/\delta))$ bits. Kane, Meka, and Nelson showed that the following approach can yield better results for $1/\varepsilon \le poly(d)$: let $\Pi = \Pi_r \times \Pi_{r-1} \times \cdots \times \Pi_1$, where Π_j has k_j-wise independent entries and maps to dimension m_j, for the k_j gradually increasing with j and the the m_j gradually decreasing (see the paper for detailed parameter settings); here $r = O(\log\log(1/\delta))$ is some parameter chosen to optimize the analysis. See Figure 1 for the final bound obtained via such an approach. It is an open problem whether $O(\log(d/\delta))$ bits is achievable.

Another related but slightly different question is that of making the JL lemma fully deterministic: given a set X of n points and a target distortion $1 + \varepsilon$, how quickly can one *deterministically* find a map $f : X \to \mathbb{R}^m$ for $m = O(\varepsilon^{-2}\log n)$ such that f has distortion at most $1 + \varepsilon$? It was shown by Engebretsen, Indyk, and O'Donnell that in fact this task can be solved in polynomial time (other later works also showed this, e.g., by Sivakumar and by Dadush et al.).

4. Fast Embeddings

As already mentioned, the primary motivation of the dimension reduction studied in this article is to improve efficiency of algorithms. The paradigm follows a two-step recipe: (1) reduce some high-dimensional input X to some lower-dimensional representation X', and then (2) run an algorithm to solve the problem on X'. Achieving smaller target dimension m tends to make step (2) more efficient, but also of importance is the time it takes to perform step (1). In particular, we want embeddings $f(x) = \Pi x$ that can be computed quickly.

The original JL distribution obtained in [JL84], as well as those studied in several subsequent works, took Π to be a dense random matrix, e.g., orthogonal projection onto a random m-dimensional subspace, or Π having i.i.d. subgaussian entries. One downside of such constructions is that mapping $x \mapsto \Pi x$ amounts to unstructured dense matrix-vector multiplication. This means computing Πx given x naively takes $O(m \cdot \|x\|_0)$ time, where $\|x\|_0$ refers to the support size $|\{i : x_i \neq 0\}|$ of x. In the worst case, this can be as bad as $\Omega(md)$ time. A natural question then is: does there exist a JL distribution allowing $x \mapsto \Pi x$ to be computed more quickly?

The first progress on a faster JL embedding was obtained by Achlioptas, who showed that if one samples $\Pi \in \mathbb{R}^{m \times d}$ to have i.i.d. (scaled) Rademacher entries, then independently zeros out each entry with probability 2/3, this is still an (ε, δ)-JL distribution for $m = O(\varepsilon^{-2} \log(1/\delta))$, and furthermore the hidden constant in the big-Oh for m is unchanged from what is achieved by the best known analysis for m in the case that Π is completely dense. The advantage of this construction is speed: if Π^i denotes the ith column of Π, then for any fixed x, the time to compute Πx is proportional to $\sum_{i \in support(x)} \|\Pi^i\|_0$. In Achlioptas' construction, the expectation of this sum is only $m/3$, obtaining a factor-3 expected speedup.

The first asymptotic speedup, i.e., by more than a constant factor, was achieved by Ailon and Chazelle [AC09]. Their idea was the following: consider Π being a (scaled) sampling matrix S, i.e., its ith row is $e_j^\top$ for a uniformly random $j \in [d]$:

$$S = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Then for any fixed x, $\mathbb{E}\,\|(1/\sqrt{m})Sx\|_2^2 = \|x\|_2^2$. The advantage here is that $\Pi = S/\sqrt{m}$ is sparse, so that Πx can be computed quickly. The problem though with obtaining a JL distribution via this construction is variance. If x for example were $e_1 = (1, 0, \dots, 0)^\top$, having Π preserve the norm of x up to $1 + \varepsilon$ with probability $1 - \delta$ would require $m = \Omega(d/(\varepsilon^2 \delta))$. Certainly, one can easily see that achieving any nontrivial guarantee would require $m \geq d$, to be expected to sample even a single nonzero entry of x. The next idea of Ailon and Chazelle was then to leverage so-called uncertainty principles from quantum mechanics: if F is the normalized Discrete Fourier Transform (so that $F^*F = I$), it is known that Fx and x cannot both be concentrated in few coordinates (concentration here is measured as $\|x\|_\infty / \|x\|_2$, which ranges between "perfectly spread" at $1/\sqrt{d}$ and "fully concentrated" on a single coordinate at 1). For example, whereas e_1 has all its mass concentrated in one coordinate, Fe_1 has its mass perfectly spread over all d coordinates. One might then be tempted to set $\Pi = SF/\sqrt{m}$ since applying F is an isometric change of basis that spreads out concentrated vectors, though of course the problem is that one might run into the opposite problem: x might be well spread, but then Fx is concentrated on a few coordinates so that sampling has poor concentration! The final observation then is that if one performs a random "phase shift," then for any fixed x, F applied to its random phase shift is likely to be well spread. In particular, Ailon and Chazelle finish their construction by selecting uniformly random $\sigma \in \{-1, 1\}^d$ and then setting $D = diag(\sigma) \in \mathbb{R}^{d \times d}$. Their final sampled mapping is then

$$\Pi = \frac{1}{\sqrt{m}} SFD.$$

They show that the resulting sample satisfies (ε, δ)-distributional JL for $m = O(\varepsilon^{-2} \log(1/\delta) \log(d/\delta))$. One can then improve this to the optimal bound by setting the final matrix to be $\Pi'\Pi$, where Π is a dense random matrix.[3] They dubbed the final embedding matrix the *Fast Johnson-Lindenstrauss Transform (FJLT)*. The overall runtime to embed x is then bounded by $O(d \log d + m^3)$, since F can be applied to any vector in $O(d \log d)$ time via the Fast Fourier Transform. It is worth noting that the DFT was not special in their analysis of correctness, and in fact any bounded orthogonal system could be used (i.e., an orthogonal matrix Q whose entries are all $O(1/\sqrt{d})$ in magnitude). For algorithmic efficiency, one also wants that Q can be applied to any fixed vector quickly; for example, one could replace F with the Hadamard-Walsh Transform. A slew of work followed, which reduced the additive $O(m^3)$ bound.

Though the FJLT improved the runtime bound over dense constructions from $O(md)$ to $O(d \log d)$ (as long as $1/\varepsilon, \log n$ are not too large) it has the downside that it does not speed up embedding time for sparse vectors, which is important in many applications. For example, in machine learning high-dimensional data is often obtained by featurizing nongeometric data, e.g., transforming an email into a histogram indexed by some dictionary, where the ith entry is the (weighted) number of occurrences of word i in the email. One can then for example train a spam classifer on the collection of high-dimensional vectors resulting from some corpus of emails. In such a case, most vectors have very small supports, since most emails do not contain every word in the dictionary! One way to embed sparse vectors faster is to make Π sparser, since if Π has at most s nonzero entries per column, then Πx can be computed naively in time $O(s \cdot \|x\|_0)$. Following work of Weinberger et al., Dasgupta, Kumar, and Sarlós [DKS10] showed that indeed sparse Π is possible. In particular, one

[3] *The actual paper describes a slightly better approach, in which S is replaced by a sparse random matrix, so that the use of Π' can be eliminated. This leads to runtime bounds that are improved by poly$(1/\varepsilon)$ factors.*

can achieve the same $m = O(\varepsilon^{-2}\log n)$ as in the JL lemma but with $s = \tilde{O}(\varepsilon^{-1}\log^3 n)$ (the $\tilde{O}$ hides $\log(\varepsilon^{-1}\log n)$ factors). Thus $s < m$ as long as $\log^2 n$ is not too large compared to $1/\varepsilon$. After a series of works, two different constructions of Kane and Nelson [KN14], both referred to as *Sparse Johnson-Lindenstrauss Transforms (SJLT)*, showed that in fact $s = O(\varepsilon^{-1}\log n)$ is achievable, which is an asymptotic improvement over dense random matrices for the full range of parameters. One construction is very simple to describe: let Π have i.i.d. (scaled) Rademacher entries and then afterward zero out all but *exactly* s nonzero entries per column. The distinction between this construction and that of Achlioptas is subtle but necessary to achieve improved bounds (recall that Achlioptas' construction sets the entries of Π to be zero independently, whereas here we enforce that number set to zero per column is fixed). The second SJLT construction is a matrix that had already been used in the streaming literature, to solve the so-called *heavy hitters* problem of finding frequent items in a data stream. In particular, it is the CountSketch matrix of Charikar, Chen, Farach-Colton. In this construction, one partitions the rows into s blocks each of equal size m/s. Then for each block of m/s rows of each column independently, we set one entry of that block, chosen uniformly at random, to be a (scaled) Rademacher. The rest of the block is set to all zero. It has also since been shown that for $m = O(\varepsilon^{-2}\log n)$, there are point sets which require columnwise sparsity s to be $\Omega((\varepsilon^{-1}/\log(1/\varepsilon))\log n)$ for any construction Π so that the SJLT achieves nearly optimal sparsity.

5. A Few Applications

In this section we highlight a few applications in which one can either use the JL lemma black box, or use more properties of random projections, to achieve good results for various applications.

5.1. Compressed sensing.

In compressed sensing [CRT06, Don06], there is a high-dimensional signal $x \in \mathbb{R}^n$ which is "compressible," i.e., approximately sparse in some known basis D. That is, there is a sparse vector w such that $\|x - Dw\|$ is small for some norm $\|\cdot\|$. The goal is to approximately recover w (and hence x) given few linear measurements. Organizing these linear measurements as rows of a matrix $A \in \mathbb{R}^{m\times n}$, we can thus rephrase by saying we would like to design an algorithm such that, given $y = Ax$, we can recover $\tilde{x}$ such that $\|w - \tilde{w}\|_Z$ is small (possibly for some other norm $\|\cdot\|_Z$ different from $\|\cdot\|$).

Candès and Tao introduced the concept of the *restricted isometry property (RIP)*, and a matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the (k,ε)-RIP if for all k-sparse signals w,

$$(1-\varepsilon)\|w\|_2^2 \le \|Aw\|_2^2 \le (1+\varepsilon)\|w\|_2^2.$$

Quantitatively improving upon previous bounds, Candès showed that if $A = \Pi D$ satisfies the $(2k, \sqrt{2}-1)$-RIP, then an optimal solution $\tilde{w}$ to the *basis pursuit* linear program

$$\begin{aligned} &\min \|\Pi D z\|_1 \\ &\text{such that } \Pi D z = \Pi D w \end{aligned}$$

is guaranteed to satisfy

$$\|w - \tilde{w}\|_1 \le \frac{1}{\sqrt{k}}\|w_{tail(k)}\|_2,$$

where $w_{tail(k)}$ is w but with top k entries in magnitude zeroed out (and hence $w_{tail(k)} = 0$ if w is actually k-sparse). It is also known that for ΠD to satisfy the (k,ε)-RIP, it is enough for the map $x \mapsto \Pi D x$ to have distortion $1 + O(\varepsilon)$ on a set of size $\binom{n}{k}\cdot\exp(O(k))$ [BDDW08, FR13], and thus such a matrix can have $O(\varepsilon^{-2}k\log(n/k))$ rows. Also, just as the JL lemma can be so-used to obtain an RIP matrix, the reverse is also true: Krahmer and Ward [KW11] showed that any $(O(\log(1/\delta)), O(\varepsilon))$-RIP matrix Π gives rise to an (ε,δ)-JL distribution defined by picking independent Rademachers $\sigma_1, \dots, \sigma_n$ and then producing the matrix $\Pi\cdot diag(\sigma)$. The best known improvements of the FJLT combine this result with analyses of RIP for Π that support fast matrix-vector multiplication, such as sampling rows from the Discrete Fourier Transform [HR17].

5.2. Randomized linear algebra.

Random projections have found a number of uses in approximation algorithms for various computational linear algebra problems as well, such as least squares regression, low-rank approximation, approximating leverage scores, distributed principle component analysis, k-means clustering, canonical correlation analysis, ℓ_p regression, ridge regression, CUR matrix factorization, and streaming approximation of eigenvalues, to name a few; see the book [Woo14].

The first work to show the connection between the JL lemma and fast approximate algorithms for linear algebra tasks is that of Papadimitriou et al., which proposed an approach to compute a low-rank (say rank k) approximation for a matrix A by computing $B = \Pi A$, then computing the singular value decomposition of B, and then projecting the columns of A onto the the subspace spanned by the top $2k$ right singular vectors of B. Later, Sarlós gave random projection-based methods for approximate least squares regression and low-rank approximation with better error guarantees. A method of Sarlós that is simple to describe is the "Sketch-and-Solve" paradigm for least squares regression: given a tall, skinny matrix $X \in \mathbb{R}^{n\times d}$ and vector $y \in \mathbb{R}^n$, to compute

$$\beta^{LS} = \underset{\beta\in\mathbb{R}^d}{argmin}\,\|X\beta - y\|_2^2,$$

we pick a random matrix $\Pi \in \mathbb{R}^{m\times n}$, $m \ll n$, from an

appropriate distribution and then compute

$$\tilde{\beta}^{LS} = \underset{\beta \in \mathbb{R}^d}{argmin} \, \|\Pi X\beta - \Pi y\|_2^2.$$

An exact solution to β^{LS} is given by $\beta^{LS} = (X^\top X)^\dagger X^T y$, where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse. The bottleneck here is computing $X^\top X$, which straightforwardly can be done in $\Theta(nd^2)$ floating point operations (flops). Meanwhile, $\tilde{\beta}^{LS} = ((\Pi X)^\top \Pi X)^\dagger (\Pi X)^\top y$ takes time equal to the sum of two terms: (1) the time to compute ΠX and Πy, given Π, X, y, and (2) the time to compute $\tilde{\beta}^{LS}$ given $\Pi X, \Pi y$. The latter requires only $\Theta(md^2)$ flops, which is much less than nd^2 since $m \ll n$. Note that if Π were a dense, unstructured matrix, then computing ΠX would take mnd time; we will eventually achieve $m \approx d/\varepsilon^2$ so that this is slower than just solving for β^{LS} directly to begin with! We thus speed up (1) by using Π which supports fast multiplication, such as a structured or sparse matrix. The key property of Π Sarlós needs in his analysis is that Π is an *ε-subspace embedding* for the subspace spanned by y and the columns of X.

For a linear subspace $E \subset \mathbb{R}^n$ and $\varepsilon \in (0,1)$, we say a matrix $\Pi \in \mathbb{R}^{m \times n}$ is an *ε-subspace embedding for E* if

$$\forall x \in E, \; (1-\varepsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\varepsilon)\|x\|_2^2.$$

In other words, Π provides a JL guarantee for an infinite set: the entire subspace E. It can be shown via a net argument that if $dim(E) = d$, then there is a finite set E' with $|E'| \le C^d$ for a universal constant $C > 1$ such that if Π satisfies the JL guarantee for E' with error parameter $\varepsilon/4$, then Π is an ε-subspace embedding for E (see [Woo14]). Then, since $\tilde{\beta}^{LS}$ is the minimizer for $\|\Pi X\beta - \Pi y\|_2^2$, we have

$$\begin{aligned}(1-\varepsilon)\|X\tilde{\beta}^{LS} - y\|_2^2 &\le \|\Pi X\tilde{\beta}^{LS} - \Pi y\|_2^2 \\ &\le \|\Pi X\beta^{LS} - \Pi y\|_2^2 \\ &\le (1+\varepsilon)\|X\beta^{LS} - y\|_2^2.\end{aligned}$$

The first and last inequalities hold using the subspace embedding guarantee, since $\Pi X\beta - \Pi y = \Pi(X\beta - y)$ and thus $X\beta - y$ is in the subspace spanned by y and the columns of X. Rearranging gives

$$\|X\tilde{\beta}^{LS} - y\|_2^2 \le \frac{1+\varepsilon}{1-\varepsilon} \cdot \|X\beta^{LS} - y\|_2^2,$$

so that $\tilde{\beta}^{LS}$ is a near-minimizer for the regression problem. It can be shown that the FJLT (see Section 4) with $O(\varepsilon^{-2} d \log^c(d/\varepsilon))$ rows provides the ε-subspace embedding guarantee with good probability. One can also take Π to be a sparse random matrix, having $m = O(d^2/\varepsilon^2)$ rows with $s = 1$ nonzero per column or $m = O(\varepsilon^{-2} d \log d)$ with $s = O(\varepsilon^{-1} \log d)$.

5.3. ***k*-means clustering.** For integer $k \ge 1$, in k-means clustering the input is $x_1, \dots, x_n \in \mathbb{R}^d$, and the goal is to find k "cluster centers" $y_1, \dots, y_k$ so as to minimize

$$\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - y_j\|_2^2.$$

Any choice of $y_1, \dots, y_k$ induces a Voronoi partition on the set of input points, i.e., $P_1, \dots, P_k$, where $P_j := \{i : j = argmin_t \|x_i - y_t\|_2^2\}$. One can then rephrase the problem as finding an optimal partition of the n points $\mathcal{P} = (P_1, \dots, P_k)$ so as to minimize

$$\sum_{j=1}^{k} \min_{y_j \in \mathbb{R}^d} \left(\sum_{i \in P_j} \|x_i - y_j\|_2^2 \right). \tag{4}$$

It can be shown that for a fixed $\mathcal{P}$, the optimal choice of the y_j to minimize equation (4) is to pick centroids $y_j = (1/|P_j|) \sum_{i \in P_j} x_i$. We can thus define

$$cost(\mathcal{P}) = \sum_{j=1}^{k} \sum_{i \in P_j} \|x_i - \frac{1}{|P_j|} \sum_{i' \in P_j} x_{i'}\|_2^2$$

and rephrase the problem as finding the k-partition $\mathcal{P}$ of minimum cost. By expanding the square (i.e., $\|a-b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2\langle a, b\rangle$) and rearranging terms, one has the equivalent definition

$$cost(\mathcal{P}) = \sum_{j=1}^{k} \sum_{i < i' \in P_j} \|x_i - x_{i'}\|_2^2. \tag{5}$$

It was first observed by Boutsidis et al. that, based on equation (5), k-means clustering can be rephrased as a constrained low-rank approximation problem. Specifically, for a k-partition $\mathcal{P} = (P_1, \dots, P_k)$ of $\{1, \dots, n\}$, define an $n \times k$ matrix $X_{\mathcal{P}}$ in which $(X_{\mathcal{P}})_{i,j}$ is $1/\sqrt{|P_j|}$ if $i \in P_j$ and is 0 otherwise. Then $X_{\mathcal{P}} X_{\mathcal{P}}^\top$ is a rank-k orthogonal projection, and furthermore the ith row of $X_{\mathcal{P}} X_{\mathcal{P}}^\top A$ is the centroid of the partition that the ith row of A belongs to in $\mathcal{P}$. Thus one can rewrite $cost(\mathcal{P})$ as $\|(I - X_{\mathcal{P}} X_{\mathcal{P}}^\top)A\|_F^2$ for $\|\cdot\|_F$ denoting the Frobenius (or Hilbert-Schmidt) norm. Then the k-means problem can be rewritten as solving the constrained low-rank approximation problem of computing

$$Q = \underset{Q \in \mathcal{Q}_k}{argmin} \, \|(I - Q)A\|_F^2,$$

where $\mathcal{Q}_k$ is the set of all rank-k projections that can be written as $X_{\mathcal{P}} X_{\mathcal{P}}^\top$ for some k-partion $\mathcal{P}$. Boutsidis et al. were then able to use this observation, coupled with randomized linear algebra techniques based on subspace embeddings (see Section 5.2) to show that if one uses a randomized linear embedding into dimension $O(k)$, the cost of k-means clustering can be preserved up to a constant factor arbitrarily close to 2. This was improved by Cohen et al.

to show that embedding into dimension $m = O(k/\varepsilon^2)$ suffices to preserve the k-means optimization problem up to $1+\varepsilon$. Most recently, Makarychev, Makarychev, and Razenshteyn in fact showed that $m = O(\log(k/\varepsilon)/\varepsilon^2)$ is sufficient (see also work by Becchetti et al.).

6. Static Data Structures for High-Dimensional Problems

In static data structural problems, one has an input set D of data items that are given up front to be preprocessed, and a set Q of possible later queries that may come about D. The goal is to (1) preprocess D quickly into a data structure that (2) uses little space, such that (3) later queries about D can be answered quickly. Consider the case that D is a set of high-dimensional vectors $X = \{x_1, \dots, x_n\}$ and queries are geometric, such as the problem of nearest neighbor search (given a query vector q, find the closest input vector x_i to q under some prescribed metric).

The downside of low-distortion embeddings such as the JL lemma is that, though such an embedding f can be found which has low distortion on X, there are no promises that it preserves distances from X to later query vectors q. The fact that the actual proof of the JL lemma is via oblivious randomized linear embeddings is actually helpful in this regard, since with high probability the distance from q to all points in X is preserved with high probability even though f was selected without knowing q. However, even so, there is a disadvantage that this distance preservation property is only probabilistically guaranteed, and furthermore even that probabilistic guarantee is void if a query vector q is selected based on adaptive interaction with the data structure (i.e., in exploratory data analysis, in which one might choose future queries based on the answers to previous queries, so that future queries are thus random and not independent of the randomness used by the data structure). A fix to this issue is to have a so-called *terminal embedding* f [EFN17], which guarantees that $\forall x \in X\ \forall y \in \mathbb{R}^d$,

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

The main difference between this guarantee and that of the JL lemma is that y can be an arbitrary point in space (e.g., the query point q) and need not be in X. Following works by Elkin et al. and Mahabadi et al., it was shown by Narayanan and Nelson that such a terminal guarantee can still be achieved with $m = O(\varepsilon^{-2}\log n)$, i.e., the same asymptotic bound as in the JL lemma.

References

[Ach03] Dimitris Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, J. Comput. System Sci. **66** (2003), no. 4, 671–687, DOI 10.1016/S0022-0000(03)00025-4. Special issue on PODS 2001 (Santa Barbara, CA). MR2005771

[AC09] Nir Ailon and Bernard Chazelle, *The fast Johnson-Lindenstrauss transform and approximate nearest neighbors*, SIAM J. Comput. **39** (2009), no. 1, 302–322, DOI 10.1137/060673096. MR2506527

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. **28** (2008), no. 3, 253–263, DOI 10.1007/s00365-007-9003-x. MR2453366

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59** (2006), no. 8, 1207–1223, DOI 10.1002/cpa.20124. MR2230846

[CW09] Kenneth L. Clarkson and David P. Woodruff, *Numerical linear algebra in the streaming model*, Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC), 2009, pp. 205–214. Full version at `www.cs.cmu.edu/afs/cs/user/dwoodruf/www/cw09.pdf`.

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós, *A sparse Johnson-Lindenstrauss transform*, STOC'10—Proceedings of the 2010 ACM International Symposium on Theory of Computing, ACM, New York, 2010, pp. 341–350. MR2743282

[Don06] David L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), no. 4, 1289–1306, DOI 10.1109/TIT.2006.871582. MR2241189

[EFN17] Michael Elkin, Arnold Filtser, and Ofer Neiman, *Terminal embeddings*, Theoret. Comput. Sci. **697** (2017), 1–36, DOI 10.1016/j.tcs.2017.06.021. MR3705706

[FR13] Simon Foucart and Holger Rauhut, *A mathematical introduction to compressive sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, New York, 2013. MR3100033

[HPIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani, *Approximate nearest neighbor: towards removing the curse of dimensionality*, Theory Comput. **8** (2012), 321–350, DOI 10.4086/toc.2012.v008a014. MR2948494

[HR17] Ishay Haviv and Oded Regev, *The restricted isometry property of subsampled Fourier matrices*, Geometric aspects of functional analysis, Lecture Notes in Math., vol. 2169, Springer, Cham, 2017, pp. 163–179. MR3645121

[Ind01] Piotr Indyk, *Algorithmic applications of low-distortion geometric embeddings*, 42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001), IEEE Computer Soc., Los Alamitos, CA, 2001, pp. 10–33. MR1948692

[JL84] William B. Johnson and Joram Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Conference in modern analysis and probability (New Haven, Conn., 1982), Contemp. Math., vol. 26, Amer. Math. Soc., Providence, RI, 1984, pp. 189–206, DOI 10.1090/conm/026/737400. MR737400

[JN10] William B. Johnson and Assaf Naor, *The Johnson-Lindenstrauss lemma almost characterizes Hilbert space, but not quite*, Discrete Comput. Geom. **43** (2010), no. 3, 542–553, DOI 10.1007/s00454-009-9193-z. MR2587836

[KMN11] Daniel Kane, Raghu Meka, and Jelani Nelson, *Almost optimal explicit Johnson-Lindenstrauss families*, Approximation, randomization, and combinatorial optimization,

Lecture Notes in Comput. Sci., vol. 6845, Springer, Heidelberg, 2011, pp. 628–639, DOI 10.1007/978-3-642-22935-0_53. MR2863296

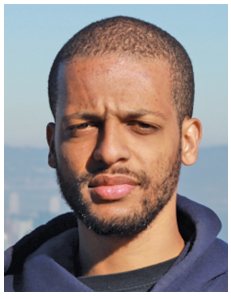
[KN14] Daniel M. Kane and Jelani Nelson, *Sparser Johnson-Lindenstrauss transforms*, J. ACM **61** (2014), no. 1, Art. 4, 23, DOI 10.1145/2559902. MR3167920

[KW11] Felix Krahmer and Rachel Ward, *New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal. **43** (2011), no. 3, 1269–1281, DOI 10.1137/100810447. MR2821584

[LN17] Kasper Green Larsen and Jelani Nelson, *Optimality of the Johnson-Lindenstrauss lemma*, 58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017, IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 633–638. MR3734267

[Vad11] Salil P. Vadhan, *Pseudorandomness*, Found. Trends Theor. Comput. Sci. **7** (2011), no. 1-3, front matter, 1–336, DOI 10.1561/0400000010. MR3019182

[Woo14] David P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci. **10** (2014), no. 1-2, iv+157, DOI 10.1561/0400000060. MR3285427



Jelani Nelson

Credits

Opening image is courtesy of the-lightwriter via Getty.
Photo of Jelani Nelson is courtesy of Yaphet Teklu.