Lower Bounds for Oblivious Subspace Embeddings

Jelani Nelson[*] Huy L. Nguyễn[†]

August 15, 2013

Abstract

An *oblivious subspace embedding (OSE)* for some $\varepsilon, \delta \in (0, 1/3)$ and $d \le m \le n$ is a distribution $\mathcal{D}$ over $\mathbb{R}^{m \times n}$ such that for any linear subspace $W \subset \mathbb{R}^n$ of dimension d,

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}} \left(\forall x \in W,\ (1-\varepsilon)\|x\|_2 \le \|\Pi x\|_2 \le (1+\varepsilon)\|x\|_2\right) \ge 1 - \delta.$$

We prove that any OSE with $\delta < 1/3$ must have $m = \Omega((d + \log(1/\delta))/\varepsilon^2)$, which is optimal. Furthermore, if every Π in the support of $\mathcal{D}$ is sparse, having at most s non-zero entries per column, then we show tradeoff lower bounds between m and s.

1 Introduction

A *subspace embedding* for some $\varepsilon \in (0, 1/3)$ and linear subspace W is a matrix Π satisfying

$$\forall x \in W,\ (1-\varepsilon)\|x\|_2 \le \|\Pi x\|_2 \le (1+\varepsilon)\|x\|_2.$$

An *oblivious subspace embedding (OSE)* for some $\varepsilon, \delta \in (0, 1/3)$ and integers $d \le m \le n$ is a distribution $\mathcal{D}$ over $\mathbb{R}^{m \times n}$ such that for any linear subspace $W \subset \mathbb{R}^n$ of dimension d,

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}} \left(\forall x \in W,\ (1-\varepsilon)\|x\|_2 \le \|\Pi x\|_2 \le (1+\varepsilon)\|x\|_2\right) \ge 1 - \delta. \tag{1}$$

That is, for any linear subspace $W \subset \mathbb{R}^n$ of bounded dimension, a random Π drawn according to $\mathcal{D}$ is a subspace embedding for W with good probability.

OSE's were first introduced in [16] and have since been used to provide fast approximate randomized algorithms for numerical linear algebra problems such as least squares regression [4, 11, 13, 16], low rank approximation [3, 4, 13, 16], minimum margin hyperplane and

[*]Harvard University. minilek@seas.harvard.edu. This work was done while the author was a member at the Institute for Advanced Study, supported by NSF CCF-0832797 and NSF DMS-1128155.

[†]Princeton University. hlnguyen@princeton.edu. Supported in part by NSF CCF-0832797 and a Gordon Wu fellowship.

minimum enclosing ball [15], and approximating leverage scores [10]. For example, consider the least squares regression problem: given $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$, compute

$$x^* = \mathrm{argmin}_{x\in\mathbb{R}^d} \|Ax - b\|_2.$$

The optimal solution x^* is such that Ax^* is the projection of b onto the column span of A. Thus by computing the singular value decomposition (SVD) $A = U\Sigma V^T$ where $U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{d\times r}$ have orthonormal columns and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the non-zero singular values of A (here r is the rank of A), we can set $x^* = V\Sigma^{-1}U^T b$ so that $Ax^* = UU^T b$ as desired. Given that the SVD can be approximated in time $\tilde{O}(nd^{\omega-1})$[1] [6] where $\omega < 2.373\ldots$ is the exponent of square matrix multiplication [18], we can solve the least squares regression problem in this time bound.

A simple argument then shows that if one instead computes

$$\tilde{x} = \mathrm{argmin}_{x\in\mathbb{R}^d} \|\Pi Ax - \Pi b\|_2$$

for some subspace embedding Π for the $(d + 1)$-dimensional subspace spanned b and the columns of A, then $\|A\tilde{x} - b\|_2 \le (1 + O(\varepsilon))\|Ax^* - b\|_2$, i.e. $\tilde{x}$ serves as a near-optimal solution to the original regression problem. The running time then becomes $\tilde{O}(md^{\omega-1})$, which can be a large savings for $m \ll n$, plus the time to compute ΠA and Πb and the time to find Π.

It is known that a random gaussian matrix with $m = O((d + \log(1/\delta))/\varepsilon^2)$ is an OSE (see for example the net argument in Clarkson and Woodruff [4] based on the Johnson-Lindenstrauss lemma and a net in [2]). While this leads to small m, and furthermore Π is oblivious to A, b so that its computation is "for free", the time to compute ΠA is $\tilde{O}(mnd^{\omega-2})$, which is worse than solving the original least squares regression problem. Sarlós constructed an OSE $\mathcal{D}$, based on the fast Johnson-Lindenstrauss transform of Ailon and Chazelle [1], with the properties that (1) $m = \tilde{O}(d/\varepsilon^2)$, and (2) for any vector $y \in \mathbb{R}^n$ and Π in the support of $\mathcal{D}$, Πy can be computed in time $O(n \log n)$ for any Π in the support of $\mathcal{D}$. This implies an approximate least squares regression algorithm running in time $O(nd \log n) + \tilde{O}(d^\omega/\varepsilon^2)$.

A recent line of work sought to improve the $O(nd \log n)$ term above to a quantity that depends only on the sparsity of the matrix A as opposed to its ambient dimension. The works [4, 11, 13] give an OSE with $m = O(d^2/\varepsilon^2)$ where every Π in the support of the OSE has only $s = 1$ non-zero entry per column. The work [13] also showed how to achieve $m = O(d^{1+\gamma}/\varepsilon^2), s = \mathrm{poly}(1/\gamma)/\varepsilon$ for any constant $\gamma > 0$. Using these OSE's together with other optimizations (for details see the reductions in [4]), these works imply approximate regression algorithms running in time $O(\mathrm{nnz}(A) + (d^3 \log d)/\varepsilon^2)$ (the $s = 1$ case), or $O_\gamma(\mathrm{nnz}(A)/\varepsilon + d^{\omega+\gamma}/\varepsilon^2)$ or $O_\gamma((\mathrm{nnz}(A) + d^2) \log(1/\varepsilon) + d^{\omega+\gamma})$ (the case of larger s). Interestingly the algorithm which yields the last bound only requires an OSE with distortion $(1+\varepsilon_0)$ for constant ε_0, while still approximately the least squares optimum up to $1 + \varepsilon$.

As seen above we now have several upper bounds, though our understanding of lower bounds for the OSE problem is lacking. Any subspace embedding, and thus any OSE, must have $m \ge d$ since otherwise some non-zero vector in the subspace will be in the kernel of Π

[1]We say $g = \tilde{O}(f)$ when $g = O(f \cdot \mathrm{polylog}(f))$.

and thus not have its norm preserved. Furthermore, it quite readily follows from the works [9, 12] that any OSE must have $m = \Omega(\min\{n, \log(d/\delta)/\varepsilon^2\})$ (see Corollary 5). Thus the best known lower bound to date is $m = \Omega(\min\{n, d + \varepsilon^{-2}\log(d/\delta)\})$, while the best upper bound is $m = O(\min\{n, (d + \log(1/\delta))/\varepsilon^2\})$ (the OSE supported only on the $n \times n$ identity matrix is indeed an OSE with $\varepsilon = \delta = 0$). We remark that although some problems can make use of OSE's with distortion $1 + \varepsilon_0$ for some constant ε_0 to achieve $(1+\varepsilon)$-approximation to the final problem, this is not always true (e.g. no such reduction is known for approximating leverage scores). Thus it is important to understand the required dependence on ε.

Our contribution I: We show that for any $\varepsilon, \delta \in (0, 1/3)$, any OSE with distortion $1 + \varepsilon$ and error probability δ must have $m = \Omega(\min\{n, (d + \log(1/\delta))/\varepsilon^2\})$, which is optimal.

We also make progress in understanding the tradeoff between m and s. The work [14] observed via a simple reduction to nonuniform balls and bins that any OSE with $s = 1$ must have $m = \Omega(d^2)$. Also recall the upper bound of [13] of $m = O(d^{1+\gamma}/\varepsilon^2), s = \mathrm{poly}(1/\gamma)/\varepsilon$ for any constant $\gamma > 0$.

Our contribution II: We show that for δ a fixed constant and $n > 100d^2$, any OSE with $m = o(\varepsilon^2 d^2)$ must have $s = \Omega(1/\varepsilon)$. Thus a phase transition exists between sparsity $s = 1$ and super-constant sparsity somewhere around m being d^2. We also show that for $m < d^{1+\gamma}$ and $\gamma \in ((10\log\log d)/(\alpha \log d), \alpha/4)$ and $2/(\varepsilon\gamma) < d^{1-\alpha}$, for any constant $\alpha > 0$, it must hold that $s = \Omega(\alpha/(\varepsilon\gamma))$. Thus the $s = \mathrm{poly}(1/\gamma)/\varepsilon$ dependence of [13] is correct (although our lower bound requires $m < d^{1+\gamma}$ as opposed to $m < d^{1+\gamma}/\varepsilon^2$).

Our proof in the first contribution follows Yao's minimax principle combined with concentration arguments and Cauchy's interlacing theorem. Our proof in the second contribution uses a bound for nonuniform balls and bins and the simple fact that for *any* distribution over unit vectors, two i.i.d. samples are not negatively correlated in expectation.

1.1 Notation

We let $O^{n\times d}$ denote the set of all $n \times d$ real matrices with orthonormal columns. For a linear subspace $W \subseteq \mathbb{R}^n$, we let $\mathbf{proj}_W : \mathbb{R}^n \to W$ denote the projection operator onto W. That is, if the columns of U form an orthonormal basis for W, then $\mathbf{proj}_W x = UU^T x$. We also often abbreviate "orthonormal" as o.n. In the case that A is a matrix, we let $\mathbf{proj}_A$ denote the projection operator onto the subspace spanned by the columns of A. Throughout this document, unless otherwise specified all norms $\|\cdot\|$ are $\ell_2 \to \ell_2$ operator norms in the case of matrix argument, and ℓ_2 norms for vector arguments. The norm $\|A\|_F$ denotes Frobenius norm, i.e. $(\sum_{i,j} A_{i,j}^2)^{1/2}$. For a matrix A, $\kappa(A)$ denotes the condition number of A, i.e. the ratio of the largest to smallest singular value. We use $[n]$ for integer n to denote $\{1, \ldots, n\}$. We use $A \lesssim B$ to denote $A \leq CB$ for some absolute constant C, and similarly for $A \gtrsim B$.

2 Dimension lower bound

Let $U \in O^{n\times d}$ be such that the columns of U form an o.n. basis for a d-dimensional linear subspace W. Then the condition in Eq. (1) is equivalent to all singular values of ΠU lying in the interval $[1-\varepsilon, 1+\varepsilon]$. Let $\kappa(A)$ denote the condition number of matrix A, i.e. its largest singular value divided by its smallest singular value, so that for any such U an OSE has $\kappa(\Pi U) \le 1+\varepsilon$ with probability $1-\delta$ over the randomness of Π. Thus $\mathcal{D}$ being an OSE implies the condition

$$\forall U \in O^{n\times d} \underset{\Pi\sim\mathcal{D}}{\mathbb{P}}\left(\kappa(\Pi U) > 1+\varepsilon\right) < \delta \tag{2}$$

We now show a lower bound for m in any distribution $\mathcal{D}$ satisfying Eq. (2) with $\delta < 1/3$. Our proof will use a couple lemmas. The first is quite similar to the Johnson-Lindenstrauss lemma itself. Without the appearance of the matrix D, it would follow from the the analyses in [5, 8] using Gaussian symmetry.

Theorem 1 (Hanson-Wright inequality [7])**.** *Let $g = (g_1, \ldots, g_n)$ be such that $g_i \sim \mathcal{N}(0,1)$ are independent, and let $B \in \mathbb{R}^{n\times n}$ be symmetric. Then for all $\lambda > 0$,*

$$\mathbb{P}\left(\left|g^T B g - \mathrm{tr}(B)\right| > \lambda\right) \lesssim e^{-\min\left\{\lambda^2/\|B\|_F^2, \lambda/\|B\|\right\}}.$$

Lemma 2. *Let u be a unit vector drawn at random from S^{n-1}, and let $E \subset \mathbb{R}^n$ be an m-dimensional linear subspace for some $1 \le m \le n$. Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix with smallest singular value σ_{min} and largest singular value σ_{max}. Then for any $0 < \varepsilon < 1$*

$$\underset{u}{\mathbb{P}}\left(\|\mathbf{proj}_E Du\|^2 \notin (\tilde{\sigma}^2 \pm \varepsilon\sigma_{max}^2)\cdot\frac{m}{n}\right) \lesssim e^{-\Omega(\varepsilon^2 m)}$$

for some $\sigma_{min} \le \tilde{\sigma} \le \sigma_{max}$.

Proof. Let the columns of $U \in O^{n\times m}$ span E, and let u_i denote the ith row of U. Let the singular values of D be $\sigma_1^2, \ldots, \sigma_n^2$. The random unit vector u can be generated as $g/\|g\|$ for a multivariate Gaussian g with identity covariance matrix. Then

$$\|\mathbf{proj}_E Du\| = \frac{1}{\|g\|}\cdot\|UU^T Dg\| = \frac{\|U^T Dg\|}{\|g\|}. \tag{3}$$

We have

$$\mathbb{E}\,\|U^T Dg\|^2 = \mathbb{E}\, g^T DUU^T Dg = \mathrm{tr}(DUU^T D) = \sum_{i=1}^n \sigma_i^2 \cdot \|u_i\|^2 = \tilde{\sigma}^2 \sum_i \|u_i\|^2 = \tilde{\sigma}^2 m,$$

for some $\sigma_{min}^2 \le \tilde{\sigma}^2 \le \sigma_{max}^2$. Also

$$\|DUU^T D\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n \sigma_i^2 \sigma_j^2 \left\langle u_i, u_j\right\rangle^2 \le \sigma_{max}^4 \sum_{i,j} \left\langle u_i, u_j\right\rangle^2 = \sigma_{max}^4 \sum_{i,j} m,$$

and $\|DUU^TD\| \le \|D\|^2 \cdot \|UU^T\| = \sigma_{max}^2$. Therefore by the Hanson-Wright inequality,

$$\mathbb{P}\left(\left|\|U^TDg\|^2 - \tilde{\sigma}^2 m\right| > \varepsilon\sigma_{max}^2 m\right) \lesssim e^{-\Omega(\min\{\varepsilon^2 m, \varepsilon m\})} = e^{-\Omega(\varepsilon^2 m)}.$$

Similarly $\mathbb{E}\,\|g\|^2 = n$ and $\|g\|$ is also the product of a matrix with orthonormal columns (the identity matrix), a diagonal matrix with $\sigma_{min} = \sigma_{max} = 1$ (the identity matrix), and a multivariate gaussian. The analysis above thus implies

$$\mathbb{P}\left(\left|\|g\|^2 - n\right| > \varepsilon n\right) \lesssim e^{-\Omega(\varepsilon^2 n)}.$$

Therefore with probability $1 - C(e^{-\Omega(\varepsilon^2 n)} + e^{-\Omega(\varepsilon^2 m)})$ for some constant $C > 0$,

$$\|\mathbf{proj}_E Du\|^2 = \frac{\|U^TDg\|^2}{\|g\|^2} = \frac{(\tilde{\sigma}^2 \pm \varepsilon\sigma_{max}^2)m}{(1 \pm \varepsilon)n} = \frac{(\tilde{\sigma}^2 \pm O(\varepsilon)\sigma_{max}^2)m}{n}$$

□

We also need the following lemma, which is a special case of Cauchy's interlacing theorem.

Lemma 3. *Suppose $A \in \mathbb{R}^{n\times m}$, $A' \in \mathbb{R}^{(n+1)\times m}$ such that $n+1 \le m$ and the first n rows of A, A' agree. Then the singular values of A, A' interlace. That is, if the singular values of A are $\sigma_1, \ldots, \sigma_n$ and those of A' are $\beta_1, \ldots, \beta_{n+1}$,*

$$\beta_1 \le \sigma_1 \le \beta_2 \le \sigma_2 \le \ldots \le \beta_n \le \sigma_n \le \beta_{n+1}.$$

Lastly, we need the following theorem and corollary, which follows from [9]. A similar conclusion can be obtained using [12], but requiring the assumption that $d < n^{1-\gamma}$ for some constant $\gamma > 0$.

Theorem 4. *Suppose $\mathcal{D}$ is a distribution over $\mathbb{R}^{m\times n}$ with the property that for any t vectors $x_1, \ldots, x_t \in \mathbb{R}^n$,*

$$\mathbb{P}_{\Pi\sim\mathcal{D}}(\forall i \in [t],\ (1-\varepsilon)\|x_i\| \le \|\Pi x_i\| \le (1+\varepsilon)\|x_i\|) \ge 1 - \delta.$$

Then $m \gtrsim \min\{n, \varepsilon^{-2}\log(t/\delta)\}$.

Proof. The proof uses Yao's minimax principle. That is, let $\mathcal{U}$ be an arbitrary distribution over t-tuples of vectors in S^{n-1}. Then

$$\mathbb{P}_{(x_1,\ldots,x_t)\sim\mathcal{U}}\ \mathbb{P}_{\Pi\sim\mathcal{D}}\left(\forall i \in [t],\ |\|\Pi x_i\|^2 - 1| \le \varepsilon\right) \ge 1 - \delta. \tag{4}$$

Switching the order of probabilistic quantifiers, an averaging argument implies the existence of a fixed matrix $\Pi_0 \in \mathbb{R}^{m\times n}$ so that

$$\mathbb{P}_{(x_1,\ldots,x_t)\sim\mathcal{U}}\left(\forall i \in [t],\ |\|\Pi_0 x\|^2 - 1| \le \varepsilon\right) \ge 1 - \delta. \tag{5}$$

The work [9, Theorem 9] gave a particular distribution $\mathcal{U}_{hard}$ for the case $t = 1$ so that no Π_0 can satisfy Eq. (5) unless $m \gtrsim \min\{n, \varepsilon^{-2}\log(1/\delta)\}$. In particular, it showed that the left hand side of Eq. (5) is at most $1 - e^{-O(\varepsilon^2 m+1)}$ as long as $m \le n/2$ in the case $t = 1$. For larger t, we simply let the hard distribution be $\mathcal{U}_{hard}^{\otimes t}$, i.e. the t-fold product distribution of $\mathcal{U}_{hard}$. Then the left hand side of Eq. (5) is at most $(1 - e^{-C(\varepsilon^2 m+1)})^t$. Let $\delta' = e^{-C(\varepsilon^2 m+1)}$. Thus $\mathcal{D}$ cannot satisfy the property in the hypothesis of the lemma if $(1 - \delta')^t < 1 - \delta$. We have $(1 - \delta')^t \le e^{-t\delta'}$, and furthermore $e^{-x} = 1 - \Theta(x)$ for $0 < x < 1/2$. Thus we must have $t\delta' = O(\delta)$, i.e. $e^{-C(\varepsilon^2 m+1)} = \delta' = O(\delta/t)$. Rerranging terms proves the theorem. □

Corollary 5. *Any OSE distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ must have $m = \Omega(\min\{n, \varepsilon^{-2}\log(d/\delta)\})$.*

Proof. We have that for any d-dimensional subspace $W \subset \mathbb{R}^n$, a random $\Pi \sim \mathcal{D}$ with probability $1 - \delta$ simultaneously preserves norms of all $x \in W$ up to $1 \pm \varepsilon$. Thus for any set of d vectors $x_1, \ldots, x_d \in \mathbb{R}^n$, a random such Π with probability $1 - \delta$ simultaneously preserves the norms of these vectors since it even preserves their span. The lower bound then follows by Theorem 4. □

Now we prove the main theorem of this section.

Theorem 6. *Let $\mathcal{D}$ be any OSE with $\varepsilon, \delta < 1/3$. Then $m = \Omega(\min\{n, d/\varepsilon^2\})$.*

Proof. We assume $d/\varepsilon^2 \le cn$ for some constant $c > 0$. Our proof uses Yao's minimax principle. Thus we must construct a distribution $\mathcal{U}_{hard}$ such that

$$\mathop{\mathbb{P}}_{U\sim\mathcal{U}_{hard}}(\kappa(\Pi_0 U) > 1 + \varepsilon) < \delta. \tag{6}$$

cannot hold for any $\Pi_0 \in \mathbb{R}^{m\times n}$ which does not satisfy $m = \Omega(d/\varepsilon^2)$. The particular $\mathcal{U}_{hard}$ we choose is as follows: we let the d columns of U be independently drawn uniform random vectors from the sphere, post-processed using Gram-Schmidt to be orthonormal. That is, the columns of U are an o.n. basis for a random d-dimensional linear subspace of $\mathbb{R}^n$.

Let $\Pi_0 = LDW^T$ be the singular value decomposition (SVD) of Π_0, i.e. $L \in O^{m\times n}, W \in O^{n\times n}$, and D is $n \times n$ with $D_{i,i} \ge 0$ for all $1 \le i \le m$, and all other entries of D are 0. Note that $W^T U$ is distributed identically as U, which is identically distributed as $W'U$ where W' is an $n \times n$ block diagonal matrix with two blocks. The upper-left block of W' is a random rotation $M \in O^{m\times m}$ according to Haar measure. The bottom-right block of W' is the $(n-m)\times(n-m)$ identity matrix. Thus it is equivalent to analyze the singular values of the matrix $LDW'U$. Also note that left multiplication by L does not alter singular values, and the singular values of $DW'U$ and $D'MA^TU$ are identical, where A is the $n \times m$ matrix whose columns are $e_1, \ldots, e_m$. Also D' is an $m \times m$ diagonal matrix with $D'_{i,i} = D_{i,i}$. Thus we wish to show that if m is sufficiently small, then

$$\mathop{\mathbb{P}}_{M\sim O^{m\times m}, U\sim\mathcal{U}_{hard}}\Big(\kappa(D'MA^TU) > 1 + \varepsilon\Big) > \frac{1}{3} \tag{7}$$

Henceforth in this proof we assume for the sake of contradiction that $m \le c\cdot\min\{d/\varepsilon^2, n\}$ for some small positive constant $c > 0$. Also note that we may assume by Corollary 5 that $m = \Omega(\min\{n, \varepsilon^{-2}\log(d/\delta)\})$.

Assume that with probability strictly larger than $2/3$ over the choice of U, we can find unit vectors z_1, z_2 so that $\|A^T U z_1\|/\|A^T U z_2\| > 1+\varepsilon$. Now suppose we have such z_1, z_2. Define $y_1 = A^T U z_1/\|A^T U z_1\|, y_2 = A^T U z_2/\|A^T U z_2\|$. Then a random $M \in O^{m\times m}$ has the same distribution as $M'T$, where M' is i.i.d. as M, and T can be any distribution over $O^{m\times m}$, so we write $M = M'T$. T may even depend on U, since $M'U$ will then still be independent of U and a random rotation (according to Haar measure). Let T be the $m \times m$ identity matrix with probability $1/2$, and R_{y_1,y_2} with probability $1/2$ where R_{y_1,y_2} is the reflection across the bisector of y_1, y_2 in the plane containing these two vectors, so that $R_{y_1,y_2} y_1 = y_2, R_{y_1,y_2} y_2 = y_1$. Now note that for any fixed choice of M' it must be the case that $\|D'M'y_1\| \geq \|D'M'y_2\|$ or $\|D'M'y_2\| \geq \|D'M'y_1\|$. Thus $\|D'M'Ty_1\| \geq \|D'M'Ty_2\|$ occurs with probability $1/2$ over T, and the reverse inequality occurs with probability $1/2$. Thus for this fixed U for which we found such z_1, z_2, over the randomness of M', T we have $\kappa(D'MA^T U) \geq \|D'MA^T U z_1\|/\|D'MA^T U z_2\|$ is greater than $1+\varepsilon$ with probability at least $1/2$. Since such z_1, z_2 exist with probability larger than $2/3$ over chioce of U, we have established Eq. (7). It just remains to establish the existence of such z_1, z_2.

Let the columns of U be $u^1, \ldots, u^d$, and define $\tilde{u}^i = A^T u^i$ and $\tilde{U} = A^T U$. Let U_{-d} be the $n \times (d-1)$ matrix whose columns are $u^1, \ldots, u^{d-1}$, and let $\tilde{U}_{-d} = A^T U_{-d}$. Write $A = A^{\|} + A^{\perp}$, where the columns of $A^{\|}$ are the projections of the columns of A onto the subspace spanned by the columns of U_{-d}, i.e. $A^{\|} = U_{-d} U_{-d}^T A$. Then

$$\|A^{\|}\|_F^2 = \|U_{-d} U_{-d}^T A\|_F^2 = \|\tilde{U}_{-d}\|_F^2 = \sum_{i=1}^{d-1} \sum_{r=1}^{m} (u_r^i)^2. \tag{8}$$

By Lemma 2 with $D = I$ and $E = \mathrm{span}(e_1, \ldots, e_m)$, followed by a union bound over the $d-1$ columns of U_{-d}, the right hand side of Eq. (8) is between $(1 - C_1\varepsilon)(d-1)m/n$ and $(1 + C_1\varepsilon)(d-1)m/n$ with probability at least $1 - C(d-1) \cdot e^{-C'C_1\varepsilon^2 m}$ over the choice of U. This is $1 - d^{-\Omega(1)}$ for $C_1 > 0$ sufficiently large since $m = \Omega(\varepsilon^{-2}\log d)$. Now, if $\kappa(\tilde{U}) > 1 + \varepsilon$ then z_1, z_2 with the desired properties exist. Suppose for the sake of contradiction that both $\kappa(\tilde{U}) \leq 1 + \varepsilon$ and $(1 - C_1\varepsilon)(d-1)m/n \leq \|\tilde{U}_{-d}\|_F^2 \leq (1 + C_1\varepsilon)(d-1)m/n$. Since the squared Frobenius norm is the sum of squared singular values, and since $\kappa(\tilde{U}_{-d}) \leq \kappa(\tilde{U})$ due to Lemma 3, all the singular values of $\tilde{U}_{-d}$, and hence $A^{\|}$, are between $(1 - C_2\varepsilon)\sqrt{m/n}$ and $(1 + C_2\varepsilon)\sqrt{m/n}$. Then by the Pythagorean theorem the singular values of $A^{\perp}$ are in the interval $[\sqrt{1 - (1 + C_2\varepsilon)^2 m/n}, \sqrt{1 - (1 - C_2\varepsilon)^2 m/n}] \subseteq [1 - (1 + C_3\varepsilon)m/n, 1 - (1 - C_3\varepsilon)m/n]$.

Since the singular values of $\tilde{U}$ and $\tilde{U}^T$ are the same, it suffices to show $\kappa(\tilde{U}^T) > 1 + \varepsilon$. For this we exhibit two unit vectors x_1, x_2 with $\|\tilde{U}^T x_1\|/\|\tilde{U}^T x_2\| > 1 + \varepsilon$. Let $B \in O^{m \times d-1}$ have columns forming an o.n. basis for the column span of $AA^T U_{-d}$. Since B has o.n. columns and u^d is orthogonal to the column span of U_{-d},

$$\|\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d\| = \|BB^T A^T u^d\| = \|B^T A^T u^d\| = \|B^T (A^{\perp})^T u^d\|.$$

Let $(A^{\perp})^T = C \Lambda E^T$ be the SVD, where $C \in \mathbb{R}^{m \times m}, \Lambda \in \mathbb{R}^{m \times m}, E \in \mathbb{R}^{n \times m}$. As usual C, E have o.n. columns, and Λ is diagonal with all entries in $[1 - (1 + C_3\varepsilon)m/n, 1 - (1 - C_3\varepsilon)m/n]$.

Condition on U_{-d}. The columns of E form an o.n. basis for the column space of $A^\perp$, which is some m-dimensional subspace of the $(n-d+1)$-dimensional orthogonal complement of the column space of U_{-d}. Meanwhile u^d is a uniformly random unit vector drawn from this orthogonal complement, and thus $\|E^T u_d\|^2 \in [(1-C_4\varepsilon)^2 m/(n-d+1), (1+C_4\varepsilon)^2 m/(n-d+1)] \subset [(1-C_5\varepsilon)m/n, (1+C_5\varepsilon)m/n]$ with probability $1-d^{-\Omega(1)}$ by Lemma 2 and the fact that $d \le \varepsilon n$ and $m = \Omega(\varepsilon^{-2}\log d)$. Note then also that $\|\Lambda E^T u^d\| = \|\tilde{u}^d\| = (1 \pm C_6\varepsilon)\sqrt{m/n}$ with probability $1-d^{-\Omega(1)}$ since Λ has bounded singular values.

Also note $E^T u/\|E^T u\|$ is uniformly random in S^{m-1}, and also $B^T C$ has orthonormal rows since $B^T C C^T B = B^T B = I$, and thus again by Lemma 2 with E being the row space of $B^T C$ and $D = \Lambda$, we have $\|B^T C \Lambda E^T u\| = \Theta(\|E^T u\| \cdot \sqrt{d/m}) = \Theta(\sqrt{d/n})$ with probability $1 - e^{-\Omega(d)}$.

We first note that by Lemma 3 and our assumption on the singular values of $\tilde{U}_{-d}$, $\tilde{U}^T$ has smallest singular value at most $(1+C_2\varepsilon)\sqrt{m/n}$. We then set x_2 to be a unit vector such that $\|\tilde{U}^T x_2\| \le (1+C_2\varepsilon)\sqrt{m/n}$.

It just remains to construct x_1 so that $\|\tilde{U}^T x_1\| > (1+\varepsilon)(1+C_2\varepsilon)\sqrt{m/n}$. To construct x_1 we split into two cases:

Case 1 ($m \le cd/\varepsilon$): In this case we choose

$$x_1 = \frac{\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d}{\|\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d\|}.$$

Then

$$\begin{aligned}
\|\tilde{U}^T x_1\|^2 &= \|\tilde{U}_{-d}^T x_1\|^2 + \left\langle \tilde{u}^d, x_1 \right\rangle^2 \\
&\ge (1-C_2\varepsilon)^2 \frac{m}{n} + \|\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d\|^2 \\
&\ge (1-C_2\varepsilon)^2 \frac{m}{n} + C\frac{d}{n}. \\
&\ge \frac{m}{n}\left((1-C_2\varepsilon)^2 + \frac{C}{c}\varepsilon\right)
\end{aligned}$$

For c small, the above is bigger than $(1+\varepsilon)^2(1+C_2\varepsilon)^2 m/n$ as desired.

Case 2 ($cd/\varepsilon \le m \le cd/\varepsilon^2$): In this case we choose

$$x_1 = \frac{1}{\sqrt{2}}\left[\overbrace{\frac{\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d}{\|\mathbf{proj}_{\tilde{U}_{-d}} \tilde{u}^d\|}}^{x^\parallel} + \overbrace{\frac{\mathbf{proj}_{\tilde{U}_{-d}^\perp} \tilde{u}^d}{\|\mathbf{proj}_{\tilde{U}_{-d}^\perp} \tilde{u}^d\|}}^{x^\perp}\right].$$

Then

$$
\begin{aligned}
\|\tilde{U}^T x_1\|^2 &= \frac{1}{2}\left\|\tilde{U}^T\left(\frac{x^{\|}}{\|x^{\|}\|}+\frac{x^{\perp}}{\|x^{\perp}\|}\right)\right\|^2 \\
&= \frac{1}{2}\left\|\tilde{U}^T_{-d}\cdot\frac{x^{\|}}{\|x^{\|}\|}\right\|^2+\frac{1}{2}\left\langle \tilde{u}^d, \frac{x^{\|}}{\|x^{\|}\|}+\frac{x^{\perp}}{\|x^{\perp}\|}\right\rangle^2 \\
&= \frac{1}{2}\left\|\tilde{U}^T_{-d}\cdot\frac{x^{\|}}{\|x^{\|}\|}\right\|^2+\frac{1}{2}\left(\|x^{\|}\|+\|x^{\perp}\|\right)^2 \\
&\geq \frac{1}{2}(1-C_2\varepsilon)^2\frac{m}{n}+\frac{1}{2}\left(\sqrt{C_4\frac{d}{n}}+\left((1-C_6\varepsilon)^2\frac{m}{n}-C_4\frac{d}{n}\right)^{1/2}\right)^2 \\
&\geq \frac{1}{2}(1-C_2\varepsilon)^2\frac{m}{n}+\frac{1}{2}\left(\sqrt{C_4\frac{d}{n}}+\left((1-C_7\varepsilon)^2\frac{m}{n}\right)^{1/2}\right)^2 \qquad (9)\\
&\geq (1-C_8\varepsilon)\frac{m}{n}+C_9\frac{\sqrt{md}}{n} \qquad (10)
\end{aligned}
$$

where Eq. (9) used that $m > cd/\varepsilon$. Now note that for $m < cd/\varepsilon^2$, the right hand side of Eq. (10) is at least $(1+10(C_2+1)\varepsilon)^2 m/n$ and thus $\|\tilde{U}^T x_1\| \geq (1+10(C_2+1)\varepsilon)\sqrt{m/n}$. $\square$

3 Sparsity Lower Bound

In this section, we consider the trade-off between m, the number of columns of the embedding matrix Π, and s, the number of non-zeroes per column of Π. In this section, we only consider the case $n \geq 100d^2$. By Yao's minimax principle, we only need to argue about the performance of a fixed matrix Π over a distribution over U. Let the distribution of the columns of U be d i.i.d. random standard basis vectors in $\mathbb{R}^n$. With probability at least 99/100, the columns of U are distinct and form a valid orthonormal basis for a d dimensional subspace of $\mathbb{R}^n$. If Π succeeds on this distribution of U conditioned on the fact that the columns of U are orthonormal with probability at least 99/100, then it succeeds in the original distribution with probability at least 98/100. In section 3.1, we show a lower bound on s in terms of ε, whenever the number of columns m is much smaller than $\varepsilon^2 d^2$. In section 3.2, we show a lower bound on s in terms of m, for a fixed $\varepsilon = 1/2$. Finally, in section 3.3, we show a lower bound on s in terms of both ε and m, when they are both sufficiently small.

3.1 Lower bound in terms of ε

Theorem 7. *If $n \geq 100d^2$ and $m \leq \varepsilon^2 d(d-1)/32$, then $s = \Omega(1/\varepsilon)$.*

Proof. We first need a few simple lemmas.

Lemma 8. *Let $\mathcal{P}$ be a distribution over vectors of norm at most 1 and u and v be independent samples from $\mathcal{P}$. Then $\mathbb{E}\langle u, v\rangle \geq 0$.*

Proof. Let $\delta = \mathbb{E}\langle u, v\rangle$. Assume for the sake of contradiction that $\delta < 0$. Take t samples $u_1, \ldots, u_t$ from $\mathcal{P}$. By linearity of expectation, we have $0 \leq \mathbb{E}(\sum_i u_i)^2 \leq t + t(t-1)\delta$. This is a contradiction because the RHS tends to $-\infty$ as $t \to \infty$. $\square$

Lemma 9. *Let X be a random variable bounded by 1 and $\mathbb{E} X \geq 0$. Then for any $0 < \delta < 1$, we have $\mathbb{P}(X \leq -\delta) \leq 1/(1+\delta)$.*

Proof. We prove the contrapositive. If $\mathbb{P}(X \leq -\delta) > 1/(1+\delta)$, then

$$\mathbb{E} X \leq -\delta\, \mathbb{P}(X \leq -\delta) + \mathbb{P}(X > -\delta) < -\delta/(1+\delta) + 1 - 1/(1+\delta) = 0.$$

$\square$

Let u_i be the i column of ΠU, r_i and z_i be the index and the value of the coordinate of the maximum absolute value of u_i, and v_i be u_i with the coordinate at position r_i removed. Let p_{2j-1}(respectively, p_{2j}) be the fractions columns of Π whose entry of maximum absolute value is on row j and is positive (respectively, negative). Let $C_{i,j}$ be the indicator variable indicating whether $r_i = r_j$ and z_i and z_j are of the same sign. Let $E = \mathbb{E}\, C_{1,2} = \sum_{i=1}^{2m} p_i^2$. Let $C = \sum_{i<j\leq d} C_{i,j}$. We have

$$\mathbb{E}\, C = \frac{d(d-1)}{2} \sum_{i=1}^{2m} p_i^2 \geq \frac{d(d-1)}{4m} \geq 8\varepsilon^{-2}$$

If i_1, i_2, i_3, i_4 are distinct then C_{i_1,i_2}, C_{i_3,i_4} are independent. If the pairs (i_1, i_2) and (i_3, i_4) share one index then $\mathbb{P}(C_{i_1,i_2} = 1 \wedge C_{i_3,i_4} = 1) = \sum_i p_i^3$ and $\mathbb{P}(C_{i_1,i_2} = 1 \wedge C_{i_3,i_4} = 0) = \sum_i p_i^2(1-p_i)$. Thus for this case,

$$\begin{aligned}
\mathbb{E}(C_{i_1,i_2} - E])(C_{i_3,i_4} - E]) &= (1 - 2\sum_i p_i^2 + \sum_i p_i^3)E^2 - 2(1-E)E\sum_i p_i^2(1-p_i) + (1-E)^2\sum_i p_i^3 \\
&= E^2 - 2E^3 + E^2\sum_i p_i^3 - (2E - 2E^2)(E - \sum_i p_i^3) + (1 - 2E + E^2)\sum_i p_i^3 \\
&= \sum_i p_i^3 - E^2 \leq \left(\sum_i p_i^2\right)^{3/2}
\end{aligned}$$

The last inequality follows from the fact that the ℓ_3 norm of a vector is smaller than its ℓ_2 norm. We have

$$\mathrm{Var}[C] = \frac{d(d-1)}{2}\mathrm{Var}[C_{1,2}] + d(d-1)(d-2)\,\mathbb{E}(C_{i_1,i_2} - \mathbb{E}\, C_{i_1,i_2})(C_{i_1,i_3} - \mathbb{E}\, C_{i_1,i_3}) \leq 4(\mathbb{E}\, C)^{3/2}.$$

Therefore,

$$\mathbb{P}(C \leq (\mathbb{E}\, C)/2) \leq \frac{4\,\mathrm{Var}[C]}{(\mathbb{E}\, C)^2} \leq O\left(\sqrt{\frac{m}{d(d-1)}}\right).$$

Thus, with probability at least $1-O(\varepsilon)$, we have $C \geq 4\varepsilon^{-2}$. We now argue that there exist $1/\varepsilon$ pairwise-disjoint pairs (a_i, b_i) such that $r_{a_i} = r_{b_i}$ and z_{a_i} and z_{b_i} are of the same sign. Indeed, let d_{2j-1} (respectively, d_{2j}) be the number of u_i's with $r_i = j$ and z_i being positive (respectively, negative). Wlog, assume that $d_1, \ldots, d_t$ are all the d_i's that are at least 2. We can always get at least $\sum_{i=1}^{t}(d_i-1)/2$ disjoint pairs. We have

$$\sum_{i=1}^{t}(d_i-1)/2 \geq \frac{1}{2}\left(\sum_{i=1}^{t} d_i(d_i-1)/2\right)^{1/2} = \frac{C^{1/2}}{2} \geq \varepsilon^{-1}$$

For each pair (a_i, b_i), by Lemmas 8 and 9, $\mathbb{P}[\langle v_{a_i}, v_{b_i}\rangle \leq -\varepsilon] \leq \frac{1}{1+\varepsilon}$ and these events for different i's are independent so with probability at least $1-(1+\varepsilon)^{-1/\varepsilon} \geq 1-e^{\varepsilon/2-1}$, there exists some i such that $\langle v_{a_i}, v_{b_i}\rangle > -\varepsilon$. For Π to be a subspace embedding for the column span of U, it must be the case, for all i, that $\|u_i\| = \|\Pi U e_i\| \geq 1-\varepsilon$. We have $|z_i| \geq s^{-1/2}\|u_i\| \geq s^{-1/2}(1-\varepsilon)\ \forall i$. Therefore, $\langle u_{a_i}, u_{b_i}\rangle \geq s^{-1}(1-\varepsilon)^2 - \varepsilon$. We have

$$\begin{aligned}\left\|\Pi U\left(\frac{1}{\sqrt{2}}(e_{a_i}+e_{b_i})\right)\right\|^2 &= \frac{1}{2}\|u_{a_i}\|^2 + \frac{1}{2}\|u_{b_i}\|^2 + \langle u_{a_i}, u_{b_i}\rangle \\ &\geq (1-\varepsilon)^2(1+s^{-1}) - \varepsilon\end{aligned}$$

However, $\|\Pi U\| \leq 1+\varepsilon$ so $s \geq (1-\varepsilon)^2/(5\varepsilon)$. ☐

3.2 Lower bound in terms of m

Theorem 10. *For $n \geq 100d^2$, $\frac{20\log\log d}{\log d} < \gamma < 1/12$ and $\varepsilon = 1/2$, if $m \leq d^{1+\gamma}$, then $s = \Omega(1/\gamma)$.*

Proof. We first prove a standard bound for a certain balls and bins problem. The proof is included for completeness.

Lemma 11. *Let α be a constant in $(0,1)$. Consider the problem of throwing d balls independently and uniformly at random at $m \leq d^{1+\gamma}$ bins with $\frac{10\log\log d}{\alpha\log d} < \gamma < 1/12$. With probability at least $99/100$, at least $d^{1-\alpha}/2$ bins have load at least $\alpha/(2\gamma)$.*

Proof. Let X_i be the indicator r.v. for bin i having $t = \alpha/(2\gamma)$ balls, and $X \stackrel{\text{def}}{=} \sum_i X_i$. Then

$$\mathbb{E}\,X_1 = \binom{d}{t} m^{-t}(1-1/m)^{d-t} \geq \left(\frac{d}{tm}\right)^t e^{-1} \geq d^{-\alpha}$$

Thus, $\mathbb{E}\,X \geq d^{1-\alpha}$. Because X_i's are negatively correlated,

$$\operatorname{Var}[X] \leq \sum_i \operatorname{Var}[X_i] = n(\mathbb{E}\,X_1 - (\mathbb{E}\,X_1)^2) \leq \mathbb{E}\,X.$$

By Chebyshev's inequality,

$$\mathbb{P}[X \leq d^{1-\alpha}/2] \leq \frac{4\operatorname{Var}[X]}{(\mathbb{E}\,X)^2} \leq 4d^{\alpha-1}$$

Thus, with probability $1-4d^{\alpha-1}$, there exist $d^{1-\alpha}/2$ bins with at least $\alpha/(2\gamma)$ balls. ☐

Next we prove a slightly weaker bound for the non-uniform version of the problem.

Lemma 12. *Consider the problem of throwing d balls independently at $m \le d^{1+\gamma}$ bins. In each throw, bin i receives the ball with probability p_i. With probability at least $99/100$, there exist $d^{1-\alpha}/2$ disjoint groups of balls of size $\alpha/(4\gamma)$ each such that all balls in the same group land in the same bin.*

Proof. The following procedure is inspired by the alias method, a constant time algorithm for sampling from a given discrete distribution (see e.g. [17]). We define a set of m virtual bins with equal probabilities of receiving a ball as follows. The following invariant is maintained: in the ith step, there are $m-i+1$ values $p_1, \ldots, p_{m-i+1}$ satisfying $\sum_j p_j = (m-i+1)/m$. In the ith step, we create the ith virtual bin as follows. Pick the smallest p_j and the largest p_k. Notice that $p_j \le 1/m \le p_k$. Form a new virtual bin from p_j and $1/m - p_j$ probability mass from p_k. Remove p_j from the collection and replace p_k with $p_k + p_j - 1/m$.

By Lemma 11, there exist $d^{1-\alpha}/2$ virtual bins receiving at least $\alpha/(2\gamma)$ balls. Since each virtual bin receives probability mass from at most 2 bins, there exist $d^{1-\alpha}/2$ groups of balls of size at least $\alpha/(4\gamma)$ such that all balls in the same group land in the same bin. □

Finally we use the above bound for balls and bins to prove the lower bound. Let p_i be the fraction of columns of Π whose coordinate of largest absolute value is on row i. By Lemma 12, there exist a row i and $\alpha/(4\gamma)$ columns of ΠU such that the coordinates of maximum absolute value of those columns all lie on row i. Π is a subspace embedding for the column span of U only if $\|\Pi U e_j\| \in [1/2, 3/2]\ \forall j$. The columns of ΠU are s sparse so for any column of ΠU, the largest absolute value of its coordinates is at least $s^{-1/2}/2$. Therefore, $\|e_i^T \Pi U\|^2 \ge \alpha/(16\gamma s)$. Because $\|\Pi U\| \le 3/2$, it must be the case that $s = \Omega(\alpha/\gamma)$.

□

3.3 Combining both types of lower bounds

Theorem 13. *For $n \ge 100d^2$, $m < d^{1+\gamma}$, $\alpha \in (0,1)$, $\frac{10 \log\log d}{\alpha \log d} < \gamma < \alpha/4$, $0 < \varepsilon < 1/2$, and $2/(\varepsilon\gamma) < d^{1-\alpha}$, we must have $s = \Omega(\alpha/(\varepsilon\gamma))$.*

Proof. Let u_i be the i column of ΠU, r_i and z_i be the index and the value of the coordinate of the maximum absolute value of u_i, and v_i be u_i with the coordinate at position r_i removed. Fix $t = \alpha/(4\gamma)$. Let p_{2i-1} (respectively, p_{2i}) be the fractions of columns of Π whose largest entry is on row i and positive (respectively, negative). By Lemma 12, there exist $d^{1-\alpha}/2$ disjoint groups of t columns of ΠU such that the columns in the same group have the entries with maximum absolute values on the same row. Consider one such group $G = \{u_{i_1}, \ldots, u_{i_t}\}$. By Lemma 8 and linearity of expectation, $\mathbb{E}\sum_{u_i,u_j \in G, i\ne j} \langle v_i, v_j\rangle \ge 0$. Furthermore, $\sum_{u_i,u_j\in G, i\ne j}\langle v_i, v_j\rangle \le t(t-1)$. Thus, by Lemma 9, $\mathbb{P}(\sum_{u_i,u_j\in G,i\ne j}\langle v_i, v_j\rangle \le -t(t-1)(\varepsilon\gamma)) \le \frac{1}{1+\varepsilon\gamma}$. This event happens independently for different groups, so with probability at least $1-(1+\varepsilon\gamma)^{-1/(\varepsilon\gamma)} \ge 1 - e^{\varepsilon\gamma/2-1}$, there exists a group G such that

$$\sum_{u_i,u_j\in G, i\ne j} \langle v_i, v_j\rangle > -t(t-1)(\varepsilon\gamma)$$

The matrix Π is a subspace embedding for the column span of U only if for all i, we have $\|u_i\| = |\Pi U e_i\| \geq (1-\varepsilon)$. We have $|z_i| \geq s^{-1/2}\|u_i\| \geq s^{-1/2}(1-\varepsilon)$. Thus, $\sum_{u_i,u_j\in G, i\neq j}\langle u_i, u_j\rangle \geq t(t-1)((1-\varepsilon)^2 s^{-1} - \varepsilon\gamma)$. We have

$$\left\|\Pi U\left(\frac{1}{\sqrt{t}}\left(\sum_{i:u_i\in G} e_i\right)\right)\right\|^2 \geq (1-\varepsilon)^2 + \frac{2}{t}\binom{t}{2}((1-\varepsilon)^2 s^{-1} - \varepsilon\gamma) \geq (1-\varepsilon)^2(1+(t-1)s^{-1}) - \alpha\varepsilon/4$$

Because $\|\Pi U\| \leq 1+\varepsilon$, we must have $s \geq \frac{(\alpha/\gamma-4)(1-\varepsilon)^2}{(16+\alpha)\varepsilon}$. $\square$

References

[1] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *Proceedings of the 10th International Workshop on Randomization and Computation (RANDOM)*, pages 272–279, 2006.

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

[4] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, 2013.

[5] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[6] James Demmel, Ioana Dumitriu, and Olga Holtz. Fast linear algebra is stable. *Numer. Math.*, 108(1):59–91, 2007.

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

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

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

[10] Michael W. Mahoney, Petros Drineas, Malik Magdon-Ismail, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[11] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, 2013.

[12] Marco Molinaro, David P. Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1738–1756, 2013.

[13] Jelani Nelson and Huy L. Nguyễn. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013.

[14] Jelani Nelson and Huy L. Nguyễn. Sparsity lower bounds for dimensionality-reducing maps. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, 2013.

[15] Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for support vector machines. In *AISTATS*, 2013.

[16] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[17] Michael D. Vose. A linear algorithm for generating random numbers with a given distribution. *Software Engineering, IEEE Transactions on*, 17(9):972–975, 1991.

[18] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 887–898, 2012.