Optimal Approximate Matrix Product in Terms of Stable Rank*

Michael B. Cohen[†1], Jelani Nelson[‡2], and David P. Woodruff[§3]

1 **MIT, Cambridge, USA**
micohen@mit.edu
2 **Harvard University, Cambridge, USA**
minilek@seas.harvard.edu
3 **IBM Research Almaden, San Jose, USA**
dpwoodru@us.ibm.com

Abstract

We prove, using the subspace embedding guarantee in a black box way, that one can achieve the spectral norm guarantee for approximate matrix multiplication with a dimensionality-reducing map having $m = O(\tilde{r}/\epsilon^2)$ rows. Here $\tilde{r}$ is the maximum stable rank, i.e., the squared ratio of Frobenius and operator norms, of the two matrices being multiplied. This is a quantitative improvement over previous work of [Magen and Zouzias, SODA, 2011] and [Kyrillidis et al., arXiv, 2014] and is also optimal for any oblivious dimensionality-reducing map. Furthermore, due to the black box reliance on the subspace embedding property in our proofs, our theorem can be applied to a much more general class of sketching matrices than what was known before, in addition to achieving better bounds. For example, one can apply our theorem to efficient subspace embeddings such as the Subsampled Randomized Hadamard Transform or sparse subspace embeddings, or even with subspace embedding constructions that may be developed in the future.

Our main theorem, via connections with spectral error matrix multiplication proven in previous work, implies quantitative improvements for approximate least squares regression and low rank approximation, and implies faster low rank approximation for popular kernels in machine learning such as the gaussian and Sobolev kernels. Our main result has also already been applied to improve dimensionality reduction guarantees for k-means clustering, and also implies new results for nonparametric regression.

Lastly, we point out that the proof of the "BSS" deterministic row-sampling result of [Batson et al., SICOMP, 2012] can be modified to obtain deterministic row-sampling for approximate matrix product in terms of the stable rank of the matrices. The original "BSS" proof was in terms of the rank rather than the stable rank.

1998 ACM Subject Classification F.2.1 Numerical Algorithms and Problems

Keywords and phrases subspace embeddings, approximate matrix multiplication, stable rank, regression, low rank approximation

Digital Object Identifier 10.4230/LIPIcs.ICALP.2016.11

* This is an extended abstract. The URL http://arxiv.org/abs/1507.02268 has our full version.
† MBC was supported by an Akamai Presidential Fellowship and NSF grant CCF-1111109.
‡ JN was supported by NSF grant IIS-1447471 and CAREER CCF-1350670, ONR grant N00014-14-1-0632 and Young Investigator N00014-15-1-2388, and a Google Faculty Research Award.
§ DPW was supported by XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory FA8750-12-C-0323.

43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016).
Editors: Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi;
Article No. 11; pp. 11:1–11:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Much recent work has successfully utilized randomized dimensionality reduction techniques to speed up solutions to linear algebra problems, with applications in machine learning, statistics, optimization, and several other domains; see the recent monographs [19, 32, 42] for more details. In our work here, we give new spectral norm guarantees for approximate matrix multiplication (AMM). Aside from AMM being interesting in its own right, it has become a useful primitive in the literature for analyzing algorithms for other large-scale linear algebra problems as well. We show applications of our new guarantees to speeding up standard algorithms for generalized regression and low-rank approximation problems. We also describe applications of our results to k-means clustering (discovered in [11]) and nonparametric regression [43].

In AMM we are given A, B each with a large number of rows n, and the goal is to compute some matrix C such that $\|C - A^TB\|_X$ is "small", for some norm $\|\cdot\|_X$. Furthermore, we would like to compute C much faster than the usual time required to exactly compute A^TB.

Work on randomized methods for AMM began with [15], which focused on $\|\cdot\|_X = \|\cdot\|_F$, i.e., Frobenius norm. They showed by picking an appropriate sampling matrix $\Pi \in \mathbb{R}^{m\times n}$, $\|(\Pi A)^T(\Pi B) - A^TB\|_F \le \varepsilon\|A\|_F\|B\|_F$ with good probability if $m = \Omega(1/\varepsilon^2)$. By a *sampling matrix*, we mean the rows of Π are independent, and each row is all zero except for a 1 in a (non-uniformly) random location. If $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times p}$, note $(\Pi A)^T(\Pi B)$ can be computed in $O(mdp)$ time once ΠA and ΠB are formed, as opposed to the straightforward $O(ndp)$ time to compute A^TB.

Frobenius error was also later achieved in [38] via a different approach, with some later optimizations in [22]. This was not via sampling, but rather to use Π drawn from a distribution satisfying an "oblivious Johnson-Lindenstrauss (JL)" guarantee, i.e. a distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ satisfying the following condition for some $\varepsilon, \delta \in (0, 1/2)$: $\forall x \in \mathbb{R}^n$, $\mathbb{P}_{\Pi\sim\mathcal{D}}(|\|\Pi x\|_2^2 - \|x\|_2^2| > \varepsilon\|x\|_2^2) < \delta$. Such a matrix Π can be taken with $m = O(\varepsilon^{-2}\log(1/\delta))$ [21]. Furthermore, one can take Π to be a Fast JL transform [1] (or any of the follow-up improvements [2, 24, 36, 4, 20]) or a sparse JL transform [14, 22] to speed up the computation of ΠA and ΠB. One could also use the Thorup-Zhang sketch [40] combined with a certain technique of [28] (see [42, Theorem 2.10] for details) to efficiently boost success probability.

Other than Frobenius error, the main other error guarantee investigated in previous work is spectral error. That is, we would like $\|C - A^TB\|$ to be small, where $\|M\|$ denotes the largest singular value of M. If one is interested in applying A^TB to some set of input vectors then this type of error is the most meaningful, since $\|C - A^TB\|$ being small is equivalent to $\|Cx\| \approx \|A^TBx\|$ for any x. The first work along these lines was again by [15], who gave a procedure based on entry-wise sampling of the entries of A and B. The works [17, 39] showed that row-sampling according to leverage scores also provides the desired guarantee with few samples.

Then [38], combined with a quantitative improvement in [9], showed that one can take a Π drawn from an oblivious JL distribution with $\delta = 2^{-\Theta(r)}$ where $r(\cdot)$ denotes rank and $r = r(A) + r(B)$. Then for Π with $m = O((r + \log(1/\delta))/\varepsilon^2)$, with probability at least $1 - \delta$ over Π,

$$\|(\Pi A)^T(\Pi B) - A^TB\| \le \varepsilon\|A\|\|B\|. \tag{1}$$

As we shall see shortly via a very simple lemma (Lemma 3), a sufficient deterministic condition implying Eq. (1) is that Π is an $O(\varepsilon)$-*subspace embedding* for the r-dimensional subspace spanned by the columns of A, B. The notion of a subspace embedding was introduced by [38].

▶ **Definition 1.** Π is an *ε-subspace embedding* for $U \in \mathbb{R}^{n\times r}$, $U^TU = I$, if Π satisfies Eq. (1) with $A = B = U$, i.e. $\|(\Pi U)^T(\Pi U) - I\| \leq \varepsilon$. This is equivalent to $\forall x \in \mathbb{R}^r,\ (1-\varepsilon)\|x\|_2^2 \leq \|\Pi U x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2$, i.e. Π preserves norms of all vectors in the subspace spanned by the columns U.

An *(ε, δ, r)-oblivious subspace embedding (OSE)* is a distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ such that $\forall U \in \mathbb{R}^{n\times r},\ U^TU = I$, it holds that $\mathbb{P}_{\Pi\sim\mathcal{D}}(\|(\Pi U)^T(\Pi U) - I\| > \varepsilon) < \delta$.

Fast subspace embeddings Π, i.e. such that the products ΠA and ΠB can be computed quickly, are known using variants on the Fast JL transform such as the Subsampled Randomized Hadamard Transform (SRHT) [38, 29, 41, 30], or via sparse subspace embeddings [9, 33, 34, 27, 12, 10]. We also refer the reader to a slightly improved analysis of the SRHT in our full version [13]. In most applications it is important to have a fast subspace embedding to shrink the time it takes to transform the input data to a lower-dimensional form. The SRHT is a randomized Π with the property that ΠA can be computed in time $O(nd\log n)$. The sparse subspace embedding constructions have some parameter m rows and exactly s non-zero entries per column, so that ΠA can be computed in time $O(s \cdot nnz(A))$, where $nnz(\cdot)$ is the number of non-zero entries, and there is a tradeoff in the upper bounds between m and s.

An issue addressed by the work of [31] is that of robustness. As stated above, achieving Eq. (1) requires Π be a subspace embedding for an r-dimensional subspace. However, consider the case when A (and similarly for B) is of high rank but can be expressed as the sum of a low-rank matrix plus high-rank noise of small magnitude, i.e., $A = \tilde{A} + E_A$ for $\tilde{A}$ of rank $r(\tilde{A}) \ll r$, and where $\|E_A\|$ is very small but E_A has high (even full) rank. One would hope the noise could be ignored, but standard results require Π to have a number of rows at least as large as r, regardless of how small the magnitude of the noise is. Another case of interest (as we will see in Section 3) is when A and B are each of high rank, but their singular values decay at some appropriate rate. As discussed in Section 3, in several applications where AMM is not the final goal but rather is used as a primitive in analyzing an algorithm for some other problem (such as k-means clustering or nonparametric regression), the matrices that arise do indeed have such decaying singular values.

The work [31] remedied this by considering the *stable ranks* $\tilde{r}(A), \tilde{r}(B)$ of A and B. Define $\tilde{r}(A) = \|A\|_F^2/\|A\|^2$. Note $\tilde{r}(A) \leq r(A)$ always, but can be much less if A has a small tail of singular values. Let $\tilde{r}$ denote $\tilde{r}(A) + \tilde{r}(B)$. Among other results, [31] showed that to achieve Eq. (1) with good probability, one can take Π to be a random (scaled) sign matrix with either $m = \Omega(\tilde{r}/\varepsilon^4)$ or $m = \Omega(\tilde{r}\log(d+p)/\varepsilon^2)$ rows. As noted in follow-up work [25], both the $1/\varepsilon^4$ dependence and the $\log(d+p)$ factor are undesirable. In their data-driven low dimensional embedding application, they wanted a dimension m independent of the original dimensions, which are assumed much larger than the stable rank, and also wanted lower dependence on $1/\varepsilon$. To this end, [25] defined the *nuclear rank* as $\tilde{nr}(A) = \|A\|_*/\|A\|$ and showed $m = \Omega(\tilde{nr}/\varepsilon^2)$ rows suffice for $\tilde{nr} = \tilde{nr}(A) + \tilde{nr}(B)$. Here $\|A\|_*$ is the nuclear norm, i.e., sum of singular values of A. Since $\|A\|_F^2$ is the sum of squared singular values, it is straightforward to see that $\tilde{nr}(A) \geq \tilde{r}(A)$ always. Thus there is a tradeoff: the stable rank guarantee is worsened to nuclear rank, but dependence on $1/\varepsilon$ is improved to quadratic.

We show switching to the weaker $\tilde{nr}$ guarantee is unnecessary by showing quadratic dependence on $1/\varepsilon$ holds even with stable rank. This answers the main open question of [31, 25].

To state our results in a more natural way, we rephrase our main result to say that we achieve

$$\|(\Pi A)^T(\Pi B) - A^T B\| \leq \varepsilon\sqrt{\left(\|A\|^2 + \frac{\|A\|_F^2}{k}\right)\left(\|B\|^2 + \frac{\|B\|_F^2}{k}\right)}. \tag{2}$$

for an arbitrary $k \geq 1$, and we do so by using subspace embeddings for $O(k)$-dimensional subspaces in a certain black box way (which will be made precise soon) regardless of the ranks of A, B.

▶ Remark 1. Note that our previously stated main contribution is equivalent, since one could set $k = \tilde{r}(A) + \tilde{r}(B)$ to arrive at the conclusion that subspace embeddings for $O(\tilde{r})$-dimensional subspaces yield the guarantee in Eq. (1). Alternatively one could obtain the Eq. (2) guarantee via Eq. (1) with error parameter $\varepsilon' = \Theta(\varepsilon \cdot \min\{1, \sqrt{(\tilde{r}(A) \cdot \tilde{r}(B))/k}\})$.

Henceforth, we use the following definition.

▶ **Definition 2.** For conforming matrices A^T, B, we say Π satisfies the (k, ε)-*approximate spectral norm matrix multiplication property ((k, ε)-AMM) for* A, B if Eq. (2) holds. If Π is random and satisfies (k, ε)-AMM with probability $1 - \delta$ for any fixed A, B, then we say Π satisfies (k, ε, δ)-AMM.

Our main contribution: We give two different characterizations for Π supporting (k, ε)-AMM, both of which imply (k, ε, δ)-AMM Π having $m = O((k + \log(1/\delta))/\varepsilon^2)$ rows. The first characterization applies to any OSE distribution for which a moment bound has been proven for $\|(\Pi U)^T(\Pi U) - I\|$ (which is true for the best analyses of all known OSE's). In this case, we show a black box theorem: any $(\varepsilon, \delta, 2k)$-OSE provides (k, ε, δ)-AMM. Since matrices with subgaussian entries and $m = \Omega((k + \log(1/\delta))/\varepsilon^2)$ are $(\varepsilon, \delta, 2k)$-OSE's, our originally stated main result follows. This result is optimal, since [35] shows any randomized distribution over Π with m rows having the (k, ε, δ)-AMM property must have $m = \Omega((k + \log(1/\delta))/\varepsilon^2)$ (the hard instance there is when $A = B = U$ has orthonormal columns, and thus rank and stable rank are equal).

Our second characterization (appearing in the full version) identifies certain deterministic conditions which, if satisfied by Π, imply the desired (k, ε)-AMM property. These conditions are of the form: (1) Π should preserve a certain set of $O(\log(1/\varepsilon))$ different subspaces of varying dimensions (all depending on k, ε and not on the ranks of A, B) with varying distortions, and (2) for a certain two matrices in our analysis, left-multiplication by Π should not increase their operator norms by more than an $O(1)$ factor. These conditions are chosen carefully so that matrices with subgaussian entries and $m = \Omega(k/\varepsilon^2)$ satisfy all conditions simultaneously with high probability, again thus proving our main result while also suggesting that the conditions we have identified are the "right" ones.

Due to the black box reliance on the subspace embedding primitive in our proofs, Π need not only be a subgaussian map. Thus not only do we improve on m compared with previous work, but also in terms of the general class of Π our result applies to. For example given our first characterization, not only does it suffice to use a random sign matrix with $\Omega(k/\varepsilon^2)$ rows, but in fact one can apply our theorem to more efficient subspace embeddings such as the SRHT or sparse subspace embeddings, or even constructions discovered in the future. That is, one can automatically transfer bounds proven for the subspace embedding property to the (k, ε)-AMM property. Thus, for example, the best known SRHT analysis (see the full version) implies (k, ε, δ)-AMM for $m = \Omega((k + \log(1/(\varepsilon\delta))\log(k/\delta))/\varepsilon^2)$ rows. For sparse subspace embeddings, the analysis in [10] implies $m = \Omega(k\log(k/\delta)/\varepsilon^2)$ suffices

with $s = O(\log(k/\delta)/\varepsilon)$ non-zeroes per column of Π. The only reason for the $\log k$ loss in m for these particular distributions is not due to our theorems, but rather due to the best analyses for the simpler *subspace embedding* property in previous work already incurring the extra $\log k$ factor (note being a subspace embedding for a k-dimensional subspace is simply a special case of (k, ε)-AMM where $A = B = U$ has k orthonormal columns). In the case of the SRHT, this extra $\log k$ factor is actually necessary [41]; for sparse subspace embeddings, it is conjectured that the $\log k$ factor can be removed and that $m = \Omega((k + \log(1/\delta))/\varepsilon^2)$ actually suffices to obtain an OSE [34, Conjecture 14]. We also discuss in Remark 2 that one can set Π to be $\Pi_1 \cdot \Pi_2$ where Π_1 has subgaussian entries with $O(k/\varepsilon^2)$ rows, and Π_2 is some other fast OSE (such as the SRHT or sparse subspace embedding), and thus one could obtain the best of both worlds: (1) Π has $O(k/\varepsilon^2)$ rows, and (2) can be applied to any $A \in \mathbb{R}^{n\times d}$ in time $T + O(km'd/\varepsilon^2)$, where T is the (fast) time to apply Π_2 to A, and m' is the number of rows of Π_2. For example, by appropriate composition as discussed in Remark 2, Π can have $O(k/\varepsilon^2)$ rows and support multiplying ΠA for $A \in \mathbb{R}^{n\times d}$ in time $O(nnz(A)) + \tilde{O}(\varepsilon^{-O(1)}(k^3 + k^2 d))$.

We also observe the proof of the main result of [3] can be modified to show that given any A, B each with n rows, and given any $\varepsilon \in (0, 1/2)$, there exists a diagonal matrix $\Pi \in \mathbb{R}^{n\times n}$ with $O(k/\varepsilon^2)$ non-zero entries, and that can be computed by a deterministic polynomial time algorithm, achieving (k, ε)-AMM. The original work of [3] achieved Eq. (1) with $m = O(r/\varepsilon^2)$ for r being the sum of ranks of A, B. The work [3] stated their result for the case $A = B$, but the general case of potentially unequal matrices reduces to this case; see Section 4. Our observation also turns out to yield a stronger form of [23, Theorem 3.3]; also see Section 4.

As mentioned, aside from AMM being interesting on its own, it is a useful primitive widely used in analyses of algorithms for several other problems, including k-means clustering [5, 11], nonparametric regression [43], linear least squares regression and low-rank approximation [38], approximating leverage scores [16], and several other problems (see [42] for a recent summary). For all these, analyses of correctness for algorithms based on dimensionality reduction via some Π rely on Π satisfying AMM for certain matrices in the analysis.

After making certain quantitative improvements to connections between AMM and applications, and combining them with our main result, in Section 3 we obtain the following new results.

1. **Generalized regression:** Given $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times p}$, consider the problem of computing $X^* = argmin_{X\in\mathbb{R}^{d\times p}} \|AX - B\|$. It is standard that $X^* = (A^TA)^+A^TB$ where $(\cdot)^+$ is the Moore-Penrose pseudoinverse. The bottleneck here is computing A^TA, taking $O(nd^2)$ time. A popular approach is to instead compute $\tilde{X} = ((\Pi A)^T(\Pi A))^+(\Pi A)^T\Pi B$, i.e., the minimizer of $\|\Pi AX - \Pi B\|$. Note that computing $(\Pi A)^T(\Pi A)$ (given ΠA) only takes a smaller $O(md^2)$ amount of time. We show that if Π satisfies $(k, O(\sqrt{\varepsilon}))$-AMM for $U_A, P_{\bar{A}}B$, and is also an $O(1)$-subspace embedding for a certain $r(A)$-dimensional subspace (see Theorem 7), then

   $$\|A\tilde{X} - B\|^2 \leq (1 + \varepsilon)\|P_AB - B\|^2 + (\varepsilon/k)\|P_AB - B\|_F^2$$

   where P_A is the orthogonal projection onto the column space of A, $P_{\bar{A}} = I - P_A$, and U_A has orthonormal columns forming a basis for the column space of A. The punchline is that if the regression error $P_{\bar{A}}B$ has high actual rank but stable rank only on the order of $r(A)$, then we obtain multiplicative spectral norm error with Π having fewer rows. Generalized regression is a natural extension of the case when B is a vector, and arises for

example in Regularized Least Squares Classification, where one has multiple (non-binary) labels, and for each label one creates a column of B; see e.g. [7] for this and variations.

2. **Low-rank approximation:** We are given $A \in \mathbb{R}^{n\times d}$ and integer $k \geq 1$, and we want to compute $A_k = argmin_{r(X)\leq k} \|A - X\|$. The Eckart-Young theorem implies A_k is obtained by truncating the SVD of A to the top k singular vectors. The standard way to use dimensionality reduction for speedup, introduced in [38], is to let $S = \Pi A$ then compute $\tilde{A} = AP_S$. Then return $\tilde{A}_k$, the best rank-k approximation of $\tilde{A}$, instead of A_k (it is known $\tilde{A}_k$ can be computed more efficiently than A_k; see [8, Lemma 4.3]). We show if Π satisfies $(k, O(\sqrt{\varepsilon}))$-AMM for U_k and $A - A_k$, and is a $(1/2)$-subspace embedding for the column space of A_k, then

$$\|\tilde{A}_k - A\|^2 \leq (1+\varepsilon)\|A - A_k\|^2 + (\varepsilon/k)\|A - A_k\|_F^2.$$

The punchline is that if the stable rank of the tail $A - A_k$ is on the same order as the rank parameter k, then standard algorithms from previous work for Frobenius multiplicative error actually in fact also provide *spectral* multiplicative error. This property indeed holds for any k for popular kernel matrices in machine learning such as the gaussian and Sobolev kernels (see [37] and Examples 2 and 3 of [43]), and low-rank approximation of kernel matrices has been applied to several machine learning problems; see [18] for a discussion.

We also explain in Section 3 how our result has already been applied in recent work on dimensionality reduction for k-means clustering [12], and how it generalizes results in [43] on dimensionality reduction for nonparametric regression to use a larger class of embeddings Π.

1.1 Preliminaries and notation

We frequently use the singular value decomposition (SVD). For a matrix $A \in \mathbb{R}^{n\times d}$ of rank r, consider the compact SVD $A = U_A \Sigma_A V_A^T$ where $U_A \in \mathbb{R}^{n\times r}$ and $V_A \in \mathbb{R}^{d\times r}$ each have orthonormal columns, and Σ_A is diagonal with strictly positive diagonal entries (the singular values of A). We assume $(\Sigma_A)_{i,i} \geq (\Sigma_A)_{j,j}$ for $i < j$. We let $P_A = U_A U_A^T$ denote the orthogonal projection operator onto the column space of A. We use $\text{span}(A)$ to refer to the subspace spanned by A's columns.

Often for a matrix A we write A_k as the best rank-k approximation to A under Frobenius or spectral error (obtained by writing the SVD of A then setting all $(\Sigma_A)_{i,i}$ to 0 for $i > k$). We often denote $A - A_k$ as $A_{\bar{k}}$. For matrices with orthonormal columns, such as U_A, $(U_A)_k$ denotes the $n \times k$ matrix formed by removing all but the first k columns of U. When A is understood from context, we often write $U\Sigma V^T$ instead of $U_A \Sigma_A V_A^T$, and U_k to denote $(U_A)_k$ (and Σ_k for $(\Sigma_A)_k$, etc.).

2 Analysis of matrix multiplication for stable rank

First we record a simple lemma relating subspace embeddings and AMM; proof in full version [13].

▶ **Lemma 3.** *Let $E = \text{span}\{A, B\}$, and let Π be an ε-subspace embedding for E. Then Eq. (1) holds.*

Lemma 3 implies that if A, B each have rank at most r, it suffices for Π to have $\Omega(r/\varepsilon^2)$ rows.

In the following subsection, we give one characterization for Π to provide (k, ε, δ)-AMM, only requiring Π to have $\Omega((k + \log(1/\delta))/\varepsilon^2)$ rows, independent of r. The other characterization also allows for this many rows, but is different in that it identifies certain deterministic conditions such that, if those hold, Π provides (k, ε)-AMM. Thus, the second characterization can even apply to deterministic Π such as the truncated SVD. We provide this second characterization only in the full version.

2.1 Characterization for (k, ε, δ)-AMM via a moment property

Here we provide a way to obtain (k, ε)-AMM for any Π whose subspace embedding property has been established using the moment method, e.g. sparse subspace embeddings [33, 34, 10], dense subgaussian matrices (as analyzed in the full version), or even the SRHT (also, as analyzed in the full version). Our approach in this subsection is inspired by the introduction of the "JL-moment property" in [22] to analyze approximate matrix multiplication with Frobenius error. The following is a generalization of [22, Definition 6.1], which was only concerned with $d = 1$.

▶ **Definition 4.** A distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ has $(\varepsilon, \delta, d, \ell)$-*OSE moments* if for all matrices $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, $\mathbb{E}_{\Pi\sim\mathcal{D}} \left\|(\Pi U)^T(\Pi U) - I\right\|^\ell < \varepsilon^\ell \cdot \delta$.

The acronym "OSE" refers to *oblivious subspace embedding*, a term coined in [34] to refer to distributions over Π yielding a subspace embedding for any fixed subspace of a particular bounded dimension with high probability. We start with a simple lemma; proof in full version.

▶ **Lemma 5.** *Suppose $\mathcal{D}$ satisfies the $(\varepsilon, \delta, 2d, \ell)$-OSE moment property and A, B (1) have the same number of rows, and (2) sum of ranks at most $2d$. Then $\mathbb{E}_{\Pi\sim\mathcal{D}} \left\|(\Pi A)^T(\Pi B) - A^T B\right\|^\ell < \varepsilon^\ell \|A\|^\ell \|B\|^\ell \delta$.*

Then, just as [22, Theorem 6.2] showed that having OSE moments with $d = 1$ implies approximate matrix multiplication with Frobenius norm error, here we show that having OSE moments for larger d implies approximate matrix multiplication with operator norm error.

▶ **Theorem 6.** *Given $k, \varepsilon, \delta \in (0, 1/2)$, let $\mathcal{D}$ be any distribution over matrices with n columns with the $(\varepsilon, \delta, 2k, \ell)$-OSE moment property for some $\ell \geq 2$. Then, for any A, B,*

$$\mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}}\left(\|(\Pi A)^T(\Pi B) - A^T B\| > \varepsilon\sqrt{(\|A\|^2 + \|A\|_F^2/k)(\|B\|^2 + \|B\|_F^2/k)}\right) < \delta \tag{3}$$

Proof. We can assume A, B each have orthogonal columns. This is since, via the full SVD, there exist orthogonal matrices R_A, R_B such that AR_A and BR_B each have orthogonal columns. Since neither left nor right multiplication by an orthogonal matrix changes operator norm,

$$\|(\Pi A)^T(\Pi B) - A^T B\| = \|(\Pi A R_A)^T(\Pi B R_B) - (AR_A)^T B R_B\|.$$

Thus, we replace A by AR_A and similarly for B. We may also assume the columns $a_1, a_2, \ldots$ of A are sorted so that $\|a_i\|_2 \geq \|a_{i+1}\|_2$ for all i. Henceforth we assume A has orthogonal columns in this sorted order (and similarly for B, with columns b_i). Now, treat A as a block matrix in which the columns are blocked into groups of size k, and similarly for B (if the number of columns of either A or B is not divisible by k, then pad them

with all-zero columns until they are). Let the spectral norm of the ith block of A be $s_i = \|a_{(i-1)\cdot k+1}\|_2$, and for B denote the spectral norm of the ith block as $t_i = \|b_{(i-1)\cdot k+1}\|_2$. These equalities for A, B hold since their columns are orthogonal and sorted by norm. We claim $\sum_i s_i^2 \leq \|A\|^2 + \|A\|_F^2/k$ (and similarly for $\sum_i t_i^2$). To see this, let the blocks of A be $A'_1, \ldots, A'_q$ where $s_i = \|A'_i\|$. Note $s_1^2 = \|A'_1\| \leq \|A\|$. Also, for $i > 1$ we have $s_i^2 = \|a_{(i-1)\cdot k+1}\|_2^2 \leq \frac{1}{k}\sum_{(i-2)\cdot k+1 \leq j \leq (i-1)\cdot k} \|a_j\|_2^2 = \frac{1}{k}\|A'_{i-1}\|_F^2$. Thus $\sum_{i>1} s_i^2 \leq \|A\|_F^2/k$.

Define $C = (\Pi A)^T(\Pi B) - A^T B$. Let $v_{\{i\}}$ denote the ith block of a vector v (the k-dimensional vector whose entries consist of entries $(i-1)\cdot k+1$ to $i \cdot k$ of v), and $C_{\{i\},\{j\}}$ the (i,j)th block of C, a $k \times k$ matrix (the entries in C contained in the ith block of rows and jth block of columns).

Now, $\|C\| = \sup_{\|x\|=\|y\|=1} x^T C y$. For any such vectors x and y, we define new vectors x' and y' whose coordinates correspond to entire blocks: we let $x'_i = \|x_{\{i\}}\|$, with y' defined analogously. We similarly define C' with entries corresponding to blocks of C, where $C'_{i,j} = \|C_{\{i\},\{j\}}\|$. Then $x^T C y \leq x'^T C' y'$, simply by bounding the contribution of each block. Thus it suffices to upper bound $\|C'\|$, which we bound by its Frobenius norm $\|C'\|_F$. Now, recalling for a random variable X that $\|X\|_\ell$ denotes $(\mathbb{E}\,|X|^\ell)^{1/\ell}$ and using Minkowski's inequality (that $\|\cdot\|_\ell$ is a norm for $\ell \geq 1$),

$$\begin{aligned}\|\|C'\|_F^2\|_{\ell/2} &= \left\|\sum_{i,j} \|(\Pi A'_i)^T(\Pi B'_j) - A'^T_i B'_j\|^2\right\|_{\ell/2} \leq \sum_{i,j} \|\|(\Pi A'_i)^T(\Pi B'_j) - A'^T_i B'_j\|^2\|_{\ell/2} \\ &\leq \sum_{i,j} \varepsilon^2 s_i^2 t_j^2 \cdot \delta^{2/\ell} \text{ (Lemma 5)} = \varepsilon^2 \left(\sum_i s_i^2\right) \cdot \left(\sum_j t_j^2\right) \delta^{2/\ell},\end{aligned}$$

which is at most $(\varepsilon\sqrt{(\|A\|^2 + \frac{\|A\|_F^2}{k})(\|B\|^2 + \frac{\|B\|_F^2}{k})}\delta^{1/\ell})^2$. Now, $\mathbb{E}\,\|C'\|_F^\ell = \|\|C'\|_F^2\|_{\ell/2}^{\ell/2}$, implying

$$\begin{aligned}&\mathbb{P}\left(\|C'\| > \varepsilon\sqrt{(\|A\|^2 + \frac{\|A\|_F^2}{k})(\|B\|^2 + \frac{\|B\|_F^2}{k})}\right) \\ &\leq \mathbb{P}\left(\|C'\|_F > \varepsilon\sqrt{(\|A\|^2 + \frac{\|A\|_F^2}{k})(\|B\|^2 + \frac{\|B\|_F^2}{k})}\right) \\ &< \frac{\mathbb{E}\,\|C'\|_F^\ell}{\left(\varepsilon\sqrt{(\|A\|^2 + \frac{\|A\|_F^2}{k})(\|B\|^2 + \frac{\|B\|_F^2}{k})}\right)^\ell},\end{aligned}$$

and the latter is at most δ. ◀

We now discuss the implications of applying Theorem 6 to specific OSE's.

2.1.1 Subgaussian maps

In the full version we show that if Π has independent subgaussian entries and $m = \Omega((k + \log(1/\delta))/\varepsilon^2)$ rows, then it satisfies the $(\varepsilon, \delta, 2k, \Theta(k + \log(1/\delta)))$ OSE moment property. Thus Theorem 6 applies to show that such Π will satisfy (k, ε, δ)-AMM.

2.1.2 SRHT

The SRHT is the matrix product $\Pi = SHD$ where $D \in \mathbb{R}^{n\times n}$ is $n \times n$ diagonal with independent ± 1 entries on the diagonal, H is a "bounded orthonormal system" (i.e. an

orthogonal matrix in $\mathbb{R}^{n\times n}$ with $\max_{i,j}|H_{i,j}| = O(1/\sqrt{n})$), and the m rows of S are independent and each samples a uniformly random element of $[n]$. Bounded orthonormal systems include the discrete Fourier matrix and the Hadamard matrix; thus such Π exist supporting matrix-vector multiplication in $O(n\log n)$ time. Thus when computing ΠA for some $n\times d$ matrix A, this takes time $O(nd\log n)$ (by applying Π to A column by column). In the full version we show that the SRHT with $m = \Omega((k+\log(1/(\varepsilon\delta))\log(k/\delta))/\varepsilon^2)$ satisfies the $(\varepsilon,\delta,2k,\log(k/\delta))$-OSE moment property, and thus provides (k,ε,δ)-AMM. Interestingly our analysis of the SRHT in the full version seems to be asymptotically tighter than any other analyses in previous work even for the basic subspace embedding property, and even slightly improves the by now standard analysis of the Fast JL transform given in [1].

2.1.3 Sparse subspace embeddings

The sparse embedding distribution with parameters m, s is as follows [9, 34, 22]. The matrix Π has m rows and n columns. The columns are independent, and for each column exactly s uniformly random entries are chosen without replacement and set to $\pm 1/\sqrt{s}$ independently; other entries in that column are set to zero. Alternatively, one could use the CountSketch [6]: the m rows are equipartitioned into s sets of size m/s each. The columns are independent, and in each column we pick exactly one row from each of the s partitions and set the corresponding entry in that column to $\pm 1/\sqrt{s}$ uniformly; the rest of the entries in the column are set to 0. Note ΠA can be multiplied in time $O(s\cdot nnz(A))$, and thus small s is desirable.

It was shown in [33, 34], slightly improving [9], that either of the above distributions satisfies the $(\varepsilon,\delta,k,2)$-OSE moment property for $m = \Omega(k^2/(\varepsilon^2\delta))$, $s = 1$, and hence (k,ε,δ)-AMM (though this particular conclusion follows easily from [22, Theorem 6.2]). It was also shown in [10], improving upon [34], that they satisfy the $(\varepsilon,\delta,k,\log(k/\delta))$-OSE moment property, and hence also (k,ε,δ)-AMM, for $m = \Omega(Bk\log(k/\delta)/\varepsilon^2), s = \Omega(\log_B(k/\delta)/\varepsilon)$ for any $B > 2$. The work [10] does not explicitly discuss the OSE moment property for sparse subspace embeddings, but it is implied; see the full version. It is conjectured that for $B = O(1)$, $m = \Omega((k+\log(1/\delta))/\varepsilon^2)$ should suffice [34, Conjecture 14].

▶ Remark 2. Currently there appears to be a tradeoff: one can either use Π s.t. ΠA can be computed quickly, such as sparse subspace embeddings or the SRHT, but then m is at least $k\log k$. Alternatively one could achieve the optimal $m = O(k/\varepsilon^2)$ using subgaussian Π, but then multiplying by Π is slower: $O(mnd)$ time for $A\in\mathbb{R}^{n\times d}$. However, settling for a tradeoff is unnecessary. One can obtain the "best of both worlds" by composition so that ΠA will have the desired $O(k/\varepsilon^2)$ rows and ΠA computed in time $O(nnz(A)) + \tilde{O}(\varepsilon^{-O(1)}(k^3+k^2d))$; see full version.

3 Applications

Spectral norm approximate matrix multiplication with dimension bounds depending on stable rank has immediate applications for the analysis of generalized regression and low-rank approximation problems. We also point out to the reader recent applications of this result to kernelized ridge regression [43] and k-means clustering [11].

3.1 Generalized regression

Here we consider generalized regression: attempting to approximate a matrix B as AX, with A of rank at most k. Let P_A be the orthogonal projection operator to the column space of A, with $P_{\bar{A}} = I - P$; then the natural best approximation will satisfy $AX = P_A B$. This minimizes

both the Frobenius and spectral norms of $AX - B$. A standard approximation algorithm for this is to replace A and B with sketches ΠA and ΠB, then solve the reduced problem exactly (see e.g. [8], Theorem 3.1). This will produce $\tilde{X} = U_A((\Pi U_A)^T \Pi U_A)^{-1}(\Pi U_A)^T \Pi B$. Below we give a lemma on the guarantees of the sketched solution in terms of properties of Π; proof is in full version.

▶ **Theorem 7.** *If* Π (1) *satisfies the* $(k, \sqrt{\varepsilon/8})$*-approximate spectral norm matrix multiplication property for* $U_A, P_{\bar{A}}B$*, and* (2) *is a* $(1/2)$*-subspace embedding for the column space of* A *(which is implied by* Π *satisfying the spectral norm approximate matrix multiplication property for* U_A *with itself), then* $\|A\tilde{X} - B\|^2 \leq (1+\varepsilon)\|P_A B - B\|^2 + (\varepsilon/k) \cdot \|P_A B - B\|_F^2$.

3.2 Low-rank approximation

Now we apply the generalized regression result from Section 3.1 to obtain a result on low-rank approximation: approximating A in the form $\tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^T$, where $\tilde{U}_k$ has only k columns and both $\tilde{U}_k$ and $\tilde{V}_k$ have orthonormal columns. Here, we consider a previous approach (see e.g. [38]): (1) let $S = \Pi A$, (2) let P_S be the orthogonal projection operator to the row space of S and $\tilde{A} = AP_S$, and (3) compute an SVD of $\tilde{A}$ and keep only the top k singular vectors, then return the resulting low rank approximation $\tilde{A}_k$ of $\tilde{A}$. It turns out computing $\tilde{A}_k$ can be done much more quickly than computing A_k; see details in [8, Lemma 4.3]. Let $A_k, U_k, A_{\bar{k}}$ be as in Section 1.1.

▶ **Theorem 8.** *If* Π (1) *satisfies the* $(k, \sqrt{\varepsilon/8})$*-approximate spectral norm matrix multiplication property for* $U_k, A_{\bar{k}}$*, and* (2) *is a* $(1/2)$*-subspace embedding for the column space of* U_k *then* $\|A - \tilde{A}_k\|^2 \leq (1+\varepsilon)\|A - A_k\|^2 + (\varepsilon/k)\|A - A_k\|_F^2$

3.3 Kernelized ridge regression

In nonparametric regression one is given data $y_i = f^*(x_i) + w_i$ for $i = 1, \ldots, n$, and the goal is to recover a good estimate for the function f^*. Here the y_i are scalars, the x_i are vectors, and the w_i are independent noise, often assumed to be distributed as mean-zero gaussian with some variance σ^2. Unlike linear regression where $f^*(x_i)$ is assumed to take the form $\langle \beta, x \rangle$ for some vector β, in nonparametric regression we allow f^* to be an arbitrary function from some function space. Naturally the goal then is to recover some $\tilde{f}$ from the data that is close to f^* whp over the noise.

Recent work [43] considers the well studied problem of obtaining $\tilde{f}$ so that $\|\tilde{f} - f^*\|_n^2$ is small with high probability over the noise w, where one uses the definition $\|f - g\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2$. The work [43] considers the case where f^* comes from a space of functions which is the closure of all functions g expressable as $g(x) = \sum_{i=1}^N \alpha_i k(x, z_i)$ over all N, $\alpha \in \mathbb{R}^N$, and vectors z_i for some PSD kernel function k. See the full version for details, but the punchline is the maximum likelihood estimator for $\tilde{f}$ is then the solution f^{LS} to a Kernelized Ridge Regression (KRR) problem, and $f^{LS}(x)$ can be expressed as a linear combination of kernel evaluations $\sum_{i=1}^n \alpha_i k(x, x_i)$. Then defining matrix K with $K_{i,j} = k(x_i, x_j)$, KRR is equivalent to computing

$$\alpha^{LS} = \underset{\alpha \in \mathbb{R}^n}{argmin} \left\{ \frac{1}{2n}\alpha^T K^2 \alpha - \frac{1}{n}\alpha^T K y + \lambda_n \alpha^T K \alpha \right\} = \left(\frac{1}{n}K^2 + 2\lambda_n K \right)^{-1} \cdot \frac{1}{n} K y,$$

which can be computed in $O(n^3)$ time. The work [43] then focuses on speeding this up, by

instead computing a solution to the lower-dimensional problem

$$\tilde{\alpha}^{LS} = \underset{\alpha\in\mathbb{R}^m}{argmin}\left\{\frac{1}{2n}\alpha^T\Pi K^2\Pi^T\alpha - \frac{1}{n}\alpha^T\Pi K y + \lambda_n\alpha^T\Pi K\Pi^T\alpha\right\}$$
$$= \left(\frac{1}{n}\Pi K^2\Pi^T + 2\lambda_n\Pi K\Pi^T\right)^{-1}\cdot\frac{1}{n}\Pi K y$$

and then returning as $\tilde{f}$ the function specified by the weight vector $\tilde{\alpha} = \Pi^T\tilde{\alpha}^{LS}$. Note that once various matrix products are formed (where the running time complexity depends on the Π being used), one only needs to invert an $m\times m$ matrix thus taking $O(m^3)$ time. They then prove that $\|\tilde{f} - f^*\|_n$ is small with high probability as long as Π satisfies two deterministic conditions (see the proof of Lemma 2 [43, Section 4.1.2], specifically equation (26) in that work): (1) Π is a $(1/2)$-subspace embedding for a particular low-dimensional subspace, and (2) $\|\Pi B\| = O(\|B\|)$ for a particular matrix B of low stable rank (B is UD_2 in [43]). Note by the triangle inequality, $\|\Pi B\| \leq \|(\Pi B)^T\Pi B - B^TB\|^{1/2} + \|B\|$, and thus it suffices for Π to provide AMM for the product B^TB, where B has low stable rank. Item (1) simply requires a subspace embedding in the standard sense, and for item (2) [43] avoided AMM by obtaining a bound on $\|\Pi B\|$ directly by their own analyses for gaussian Π and the SRHT. Our result thus provides a unifying analysis which works for a larger and general class of Π, including for example sparse subspace embeddings.

3.4 *k*-means clustering

In the works [5, 11], the authors considered dimensionality reduction methods for k-means clustering. Recall in k-means clustering one is given n points $x_1,\ldots,x_n\in\mathbb{R}^d$, as well as an integer $k\geq 1$, and the goal is to find k points $y_1,\ldots,y_k\in\mathbb{R}^d$ minimizing $\sum_{i=1}^n\min_{j=1}^k\|x_i - y_j\|_2^2$. One key observation common to both [5, 11] is that k-means clustering is closely related to the problem of low-rank approximation. More specifically, given a partition $\mathcal{P} = \{P_1,\ldots,P_k\}$, define the $n\times k$ matrix $X_{\mathcal{P}}$ by $(X_{\mathcal{P}})_{i,j}$ is $1/\sqrt{|P_j|}$ if $i\in P_j$, and zero otherwise. Let $A\in\mathbb{R}^{n\times d}$ have rows $x_1,\ldots,x_n$. Then the k-means problem can be rewritten as computing $\mathcal{P}^* = \text{argmin}_{\mathcal{P}}\|A - X_{\mathcal{P}}X_{\mathcal{P}}^TA\|_F^2$, where $\mathcal{P}$ ranges over all partitions of $\{1,\ldots,n\}$ into k sets (the y_i are the distinct rows of $X_{\mathcal{P}}X_{\mathcal{P}}^TA$). It is easy to verify the columns of $X_{\mathcal{P}}$ are orthonormal, so $X_{\mathcal{P}}X_{\mathcal{P}}^T$ is the orthogonal projection onto the column space of $X_{\mathcal{P}}$. Thus if one defines $\mathcal{S}$ as the set of all rank k orthogonal projections obtained as $X_{\mathcal{P}}X_{\mathcal{P}}^T$ for some k-partition $\mathcal{P}$, then the above can be rewritten as the *constrained rank-k projection problem* of computing $\mathcal{P}^* = \text{argmin}_{P\in\mathcal{S}}\|(I-P)A\|_F^2$.

The work [11] showed that if $\mathcal{S}$ is any subset of projections of rank at most k (henceforth *rank-k projections*) and $\Pi\in\mathbb{R}^{m\times d}$ satisfies certain technical conditions to be divulged soon, then if $\tilde{P}\in\mathcal{S}$ minimizes the Π-reduced problem $\min_{P\in\mathcal{S}}\|(I-P)A\Pi^T\|_F^2$ up to a factor of γ, then $\tilde{P}$ minimizes the original problem $\min_{P\in\mathcal{S}}\|(I-P)A\|_F^2$ up to $(1+O(\varepsilon))\gamma$.

One set of sufficient conditions for Π is as follows (see [11, Lemma 10]). There is a matrix $B\in\mathbb{R}^{(n+2k)\times d}$ of stable rank $O(k)$, where k is the number of cluster centers y_i above, such that if

$$\|(\Pi B^T)^T(\Pi B^T) - BB^T\| < \varepsilon, \tag{4}$$
$$\text{and } \left|\|\Pi B_2\|_F^2 - \|B_2\|_F^2\right| \leq \varepsilon k \tag{5}$$

then $\tilde{P}$ provides good error as discussed above. Thus for Eq. (4) it suffices for Π to provide $(O(k),\varepsilon/2)$-AMM for B^T, B^T, and our results apply. Obtaining Eq. (5) is much simpler and can be derived from the JL moment property (see the proof of [22, Theorem 6.2]).

Without our results on stable-rank AMM provided in this current work, [11] gave a different analysis, avoiding [11, Lemma 10], which for subgaussian Π required $m = \Theta(k \cdot \log(1/\delta)/\varepsilon^2)$ rows (note the product between k and $\log(1/\delta)$ instead of the sum).

4 Stable rank and row selection

As well as random projections, AMM (and subspace embeddings) by row selection are also common in algorithms. This corresponds to setting Π to a diagonal matrix S with relatively few nonzero entries. Unlike random projections, there are no *oblivious* distributions of such matrices S with universal guarantees. Instead, S must be determined (either randomly or deterministically) from the matrices being embedded.

There are two particularly algorithmically useful methods for obtaining such S. The first is importance sampling: independent random sampling of the rows, but with nonuniform sampling probabilities. For rank-k matrices, $O(k(\log k)/\varepsilon^2)$ samples suffice [17, 39]. The second method is the deterministic selection method given in [3], often called "BSS", choosing only $O(k/\varepsilon^2)$ rows. This still runs in polynomial time, but originally required many expensive linear algebra steps and thus was slower in general; see [26] for runtime improvements.

The method used in [39] (matrix Chernoff bound) can be extended to the stable-rank case, making even the log factor in the number of samples depend only on the stable rank; see the full version for details. We here give an extension of BSS that covers low stable rank matrices as well. The proof is in the full version, and follows by observing that it suffices to just perform a slight modification of the original BSS proof.

▶ **Theorem 9.** *Given two matrices A and B, each with n rows, and an $\varepsilon \in (0, 1)$, there exists a diagonal matrix S with $O(k/\varepsilon^2)$ nonzero entries satisfying the (k, ε)-AMM property for A, B. Such an S can be computed by a polynomial-time algorithm.*

When $A = B$ and $A^T A$ is the identity, this is just the original BSS result. It is also stronger than Theorem 3.3 of [23], implying it when A is the combination of the rows $\sqrt{N/T} \cdot v_i$ from that theorem statement with an extra column containing the costs, and a constant ϵ. The techniques in that paper, on the other hand, can prove a result comparable to Theorem 9, but with the row count scaling as k/ε^3 rather than k/ε^2.

Acknowledgments. We thank Jarosław Błasiok for pointing out the connection between low stable rank approximate matrix multiplication and the analyses in [43].

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 Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013.

3 Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

4 Jean Bourgain. An improved estimate in the restricted isometry problem. *Geometric Aspects of Functional Analysis*, 2116:65–70, 2014.

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

6 Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

7 Pei-Chun Chen, Kuang-Yao Lee, Tsung-Ju Lee, Yuh-Jye Lee, and Su-Yun Huang. Multiclass support vector classification via coding and regression. *Neurocomputing*, 73(7-9):1501–1512, 2010.

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

9 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)*, pages 81–90, 2013. Full version at `http://arxiv.org/abs/1207.6365v4`.

10 Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 278–287, 2016.

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

12 Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proc. of the 6th Annual Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 181–190, 2015.

13 Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *CoRR*, abs/1507.02268, 2015.

14 Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, 2010.

15 Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices I: approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.

16 Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

17 Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1127–1136, 2006.

18 Alex Gittens and Michael W. Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 567–575, 2013.

19 Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

20 Ishay Haviv and Oded Regev. The restricted isometry property of subsampled Fourier matrices. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear*, 2016.

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

22 Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4, 2014.

23 Alexandra Kolla, Yury Makarychev, Amin Saberi, and Shang-Hua Teng. Subgraph sparsification and nearly optimal ultrasparsifiers. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 57–66, 2010.

24 Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

25 Anastasios T. Kyrillidis, Michail Vlachos, and Anastasios Zouzias. Approximate matrix multiplication with application to linear embeddings. *CoRR*, abs/1403.7683, 2014.

26 Yin Tat Lee and He Sun. Constructing linear sized spectral sparsification in almost linear time. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 250–269, 2015.

27 Mu Li, Gary L. Miller, and Richard Peng. Iterative row sampling. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013.

28 Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2014.

29 Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

30 Yichao Lu, Paramveer Dhillon, Dean Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized Hadamard transform. In *Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2013.

31 Avner Magen and Anastasios Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1422–1436, 2011.

32 Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

33 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)*, pages 91–100, 2013.

34 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)*, pages 117–126, 2013.

35 Jelani Nelson and Huy L. Nguyễn. Lower bounds for oblivious subspace embeddings. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 883–894, 2014.

36 Jelani Nelson, Eric Price, and Mary Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

37 Nima Reyhani, Hideitsu Hino, and Ricardo Vigário. New probabilistic bounds on eigenvalues and eigenvectors of random kernel matrices. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 627–634, 2011.

38 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.

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

40 Mikkel Thorup and Yin Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 41(2):293–331, 2012.

41 Joel A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1–2):115–126, 2011.

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

43 Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *CoRR*, abs/1501.06195, 2015.