Optimal approximate matrix product in terms of stable rank

Michael B. Cohen* Jelani Nelson† David P. Woodruff‡

Abstract

We give two different proofs that use the subspace embedding guarantee in a black box way to show that one can achieve the spectral norm guarantee for approximate matrix multiplication with a dimensionality-reducing map that has $m = O(\tilde{r}/\varepsilon^2)$ 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 strict quantitative improvement over previous work of [MZ11, KVZ14], 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 (although for some of these constructions we lose logarithmic factors, since logarithmic factors are lost in previous work even just to achieve the simpler subspace embedding property, sometimes necessarily so).

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. We furthermore give quantitative improvements to the connections proven in previous work to achieve even better bounds. Our main result has also already been applied to improve dimensionality reduction guarantees for k-means clustering [CEM+15], and also implies new results for dimensionality reduction applied to nonparametric regression [YPW15].

We also separately point out that the proof of the "BSS" deterministic row-sampling result of [BSS12] can be modified to show that for any matrices A, B of stable rank at most $\tilde{r}$, one can achieve the spectral norm guarantee for approximate matrix multiplication of $A^T B$ using a deterministic sampling matrix with $O(\tilde{r}/\varepsilon^2)$ non-zero entries which can be found in polynomial time. The original result of [BSS12] was for rank instead of stable rank. Our observation leads to a stronger version of a main theorem of [KMST10].

1 Introduction

A recent line of research has 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 [HMT11, Mah11, Woo14] for more details.

*MIT. `micohen@mit.edu`. Work done under support of an Akamai Presidential Fellowship and NSF grant CCF-1111109.

†Harvard University. `minilek@seas.harvard.edu`. Supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR grant N00014-14-1-0632 and Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

‡IBM Almaden. `dpwoodru@us.ibm.com`. Supported by XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory FA8750-12-C-0323.

In our work here, we give new spectral norm guarantees for approximate matrix multiplication, and we show applications of these guarantees to speeding up standard algorithms for generalized regression and low-rank approximation problems, and we also describe applications of our results to k-means clustering (discovered in [CEM+15]) and nonparametric regression [YPW15].

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

Work on randomized methods for approximate matrix multiplication began with [DKM06], which focused on $\|\cdot\|_X = \|\cdot\|_F$, i.e., Frobenius norm error. They showed that by picking an appropriate sampling matrix $\Pi \in \mathbb{R}^{m\times n}$, one obtains

$$\|(\Pi A)^T(\Pi B) - A^T B\|_F \leq \varepsilon\|A\|_F\|B\|_F \tag{1}$$

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 random location according to some appropriate (non-uniform) distribution. If $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times p}$, note that $(\Pi A)^T(\Pi B)$ can be computed in $O(mdp)$ time once ΠA and ΠB are formed, as opposed to the straightforward $O(ndp)$ time algorithm to compute $A^T B$.

The Frobenius norm error guarantee of Eq. (1) was also later achieved in [Sar06, Lemma 6] via a different approach, with some later optimizations to the parameters in [KN14, Theorem 6.2]. The approach of Sarlós was not via sampling, but rather to use a matrix 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 $\varepsilon, \delta \in (0, 1/2)$:

$$\forall x \in \mathbb{R}^n,\ \mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}}\left(\left|\|\Pi x\|_2^2 - \|x\|_2^2\right| > \varepsilon\|x\|_2^2\right) < \delta. \tag{2}$$

Such a matrix Π can be taken with $m = O(\varepsilon^{-2}\log(1/\delta))$ [JL84]. Furthermore, one can take Π to be a Fast JL transform [AC09] (or any of the follow-up improvements [AL13, KW11, NPW14]) or a sparse JL transform [DKS10, KN14] to speed up the computation of ΠA and ΠB. One could also use the Thorup-Zhang sketch [TZ12] combined with a certain technique of [LBKW14] (see [Woo14, Theorem 2.10] for details) to efficiently boost success probability.

Other than Frobenius norm error, the main other type of error guarantee investigated in previous work is spectral error. That is, we would like $\|C - A^T B\|$ to be small, where $\|M\|$ denotes the largest singular value of M. If one is interested in applying $A^T B$ to some set of input vectors then this type of error is the most meaningful, since $\|C - A^T B\|$ being small is equivalent to $\|Cx\| \approx \|A^T Bx\|$ for any x. The first work along these lines was again by [DKM06], who gave a procedure based on entry-wise sampling of the entries of A and B.

Then [Sar06], combined with a quantitative improvement in [CW13], showed that one can take a Π drawn from an oblivious JL distribution with $\delta = 2^{-\Theta(r)}$ where $r(\cdot)$ denotes rank and $r = r(A) + r(B)$. In this case Π has $m = O((r + \log(1/\delta))/\varepsilon^2)$, and with probability at least $1 - \delta$ with Π drawn according to $\mathcal{D}$,

$$\|(\Pi A)^T(\Pi B) - A^T B\| \leq \varepsilon\|A\|\|B\|. \tag{3}$$

As we shall see shortly via a very simple lemma (Lemma 1), a sufficient deterministic condition implying Eq. (3) is that Π is an $O(\varepsilon)$-*subspace embedding* for the r-dimensional subspace spanned by the columns of A and B. The notion of a subspace embedding was introduced by Sarlós, and we say Π is an ε-subspace embedding for the column space of some $U \in \mathbb{R}^{n\times r}$, $U^T U = I$, if Π satisfies Eq. (3) with $A = B = U$. This is equivalent to $\forall x \in \mathbb{R}^r,\ (1-\varepsilon)\|x\|_2^2 \le \|\Pi U x\|_2^2 \le (1+\varepsilon)\|x\|_2^2$.

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 SRHT [Sar06, Tro11, LDFU13] (also see a slightly improved analysis of the SRHT over previous works in Section A.2) or via sparse subspace embeddings [CW13, MM13, NN13, LMP13, CLM$^+$15]. 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 before being processed. The SRHT is a construction of a Π such that ΠA can be computed in time $O(nd \log n)$ (see Section A.2 for details of the construction). 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 $s \cdot nnz(A)$, where $nnz(\cdot)$ is the number of non-zero entries, and there is a tradeoff in the upper bounds between m and s.

One issue addressed by the work of [MZ11] is that of robustness. As stated above, achieving the guarantee of Eq. (3) requires Π to 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 $\tilde{A}$ a matrix of rank $r(\tilde{A}) \ll r(A)$, and where $\|E_A\|$ is very small but E_A has high (even full) rank. One would hope that small noise could be ignored, but standard results require Π to have a number m of rows at least as large as the rank of A, B, 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.

The work [MZ11] 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 $\tilde{r}(A) \le r(A)$ 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, [MZ11] showed that to achieve Eq. (3) 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 [KVZ14], 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 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, [KVZ14] 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 $\tilde{nr}(A) \ge \tilde{r}(A)$ always. Thus there is a tradeoff: the stable rank guarantee is worsened to nuclear rank, but dependence on $1/\varepsilon$ is improved to quadratic.

In our work, we show that the switch to the weaker nuclear rank guarantee is unnecessary. In particular, we show that the quadratic dependence on $1/\varepsilon$ is true even with stable rank. This answers the main open question of [MZ11, KVZ14].

Our main contribution: We give two different proofs using the subspace embedding guarantee *in a black box way* to show one can achieve Eq. (3) with Π having $m = O(\tilde{r}/\varepsilon^2)$ rows. Due to the black box nature of our proofs, Π can be drawn from any subspace embedding family. This is an improvement to [MZ11, KVZ14] not only quantitatively in terms of m, but also in terms of the general class of Π it applies to. That is, not only does it suffice to use a random sign matrix with $\Omega(\tilde{r}/\varepsilon^2)$ rows, but in fact one can apply our theorem with more efficient subspace embeddings such as the SRHT or sparse subspace embeddings (albeit with logarithmic factor losses in $\tilde{r}$, since those losses are incurred in current proofs even for the weaker subspace embedding guarantee, and a logarithmic factor loss is necessary for the SRHT [Tro11]), or even with subspace embedding constructions that may be developed in the future. Our bound of $m = O(\tilde{r}/\varepsilon^2)$ is optimal up to a constant factor for any oblivious dimensionality reducing map Π, as can be seen from the lower bound in [NN14], which provides lower bounds in terms of rank, but for the matrices in the hard distribution in that paper, their rank is equal to their stable rank.

We also point out that the proof of the main result of [BSS12] can be modified to show that given any A, B each with n rows and of stable rank at most $\tilde{r}$, and given any $\varepsilon \in (0, 1/2)$, there exists a diagonal matrix $\Pi \in \mathbb{R}^{n\times n}$ with $O(\tilde{r}/\varepsilon^2)$ non-zero entries, and that can be computed by a deterministic polynomial time algorithm, achieving Eq. (3). The original work of [BSS12] proved this theorem but with the $\tilde{r}$ term replaced by the maximum rank of A, B ([BSS12] 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 [KMST10, Theorem 3.3].

Aside from approximate matrix multiplication (and the special case of subspace embeddings) being interesting in its own right, it is also applicable to several other problems, including k-means clustering [BZMD15, CEM+15], nonparametric regression [YPW15], linear least squares regression and low-rank approximation [Sar06], approximating leverage scores [DMMW12], and several other problems (see [Woo14] for a recent summary). To state some of our applications in a more natural way, we rephrase our main result to say that we achieve the error guarantee

$$\|(\Pi A)^T(\Pi B) - A^T B\| \le \varepsilon\sqrt{\left(\|A\|^2 + \frac{\|A\|_F^2}{k}\right)\left(\|B\|^2 + \frac{\|B\|_F^2}{k}\right)}. \tag{4}$$

for an arbitrary $k \ge 1$, and we do so by using subspace embeddings for $O(k)$-dimensional subspaces in a black box way (see Section 2). 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. (3). Alternatively one could obtain Eq. (4) guarantee via Eq. (3) 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 1. *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.* (4) *holds.*

After making certain quantitative improvements to some of the connections between approximate matrix multiplication 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 $X^* = (A^T A)^+ A^T B$ where $(\cdot)^+$ is the Moore-Penrose pseudoinverse. The bottleneck here is computing $A^T A$, 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 3), then

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

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. [CLL+10] for this and variations.

2. **Low-rank approximation:** In this problem we are given $A \in \mathbb{R}^{n\times d}$ and an integer $k \geq 1$, and we would like to compute $A_k = argmin_{r(X)\leq k} \|A - X\|$. It is well known that A_k can be obtained by truncating the SVD of A to contain only the top k singular vectors. The standard way to use dimensionality reduction to obtain an approximate low rank approximation, introduced in [Sar06], is to let $S = \Pi A$ then compute $\tilde{A} = AP_S$. Then one returns $\tilde{A}_k$, 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 [CW09, Lemma 4.3]). We show that 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 the standard algorithms from previous work for Frobenius norm multiplicative error actually in fact also provide *spectral* multiplicative error.

We also explain in Section 3 how our result has already been applied in recent work on dimensionality reduction for k-means clustering [CLM+15], and how our result can be used to generalize results in [YPW15] on dimensionality reduction for nonparametric regression to extend to a larger class of subspace embeddings Π, such as sparse subspace embeddings.

1.1 Preliminaries and notation

Note for conforming matrices A^T, B each of stable rank at most $\tilde{r}$, Eq. (4) with $k = \tilde{r}$ and error parameter $\varepsilon/2$ implies

$$\|(\Pi A)^T(\Pi B) - A^T B\| \leq \varepsilon\|A\|\|B\| \tag{5}$$

In other words, Eq. (4) will give us results for matrices of *stable rank k* similar to those we have for matrices of *rank* k.

Throughout the paper 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.

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

Definition 2. *Let E be a linear subspace of $\mathbb{R}^n$, and let Π be an $m \times n$ matrix. Then we say Π is an ε-*subspace embedding for E *if for all $x \in E$*

$$(1-\varepsilon)\|x\|^2 \le \|\Pi x\|^2 \le (1+\varepsilon)\|x\|^2,$$

or equivalently

$$\|(\Pi U)^T(\Pi U) - U^T U\| \le \varepsilon$$

for a matrix U whose columns form an orthonormal basis for E.

First we record a simple lemma stating that subspace embeddings provide approximate matrix multiplication with respect to $\|\cdot\|$.

Lemma 1. *Let $E = \text{span}\{\text{columns}(A), \text{columns}(B)\}$, and let Π be an ε-subspace embedding for E. Then Eq. (5) holds.*

Proof. First, without loss of generality we may assume $\|A\| = \|B\| = 1$ since we can divide both sides of Eq. (5) by $\|A\| \cdot \|B\|$. Let U be a matrix whose columns form an orthonormal basis for E. Then note for any x, y we can write $Ax = Uw, By = Uz$ where $\|w\| \le \|x\|, \|z\| \le \|y\|$. Then

$$\begin{aligned}\|(\Pi A)^T(\Pi B) - A^T B\| &= \sup_{\|x\|=\|y\|=1} |\langle \Pi Ax, \Pi By\rangle - \langle Ax, By\rangle| \\ &= \sup_{\|w\|,\|z\|\le 1} |\langle \Pi Uz, \Pi Uw\rangle - \langle Uz, Uw\rangle| \\ &= \|(\Pi U)^T(\Pi U) - I\| \\ &< \varepsilon\end{aligned}$$

■

Lemma 1 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 two subsections, we give two different analyses showing Eq. (4) can be achieved with Π only having $\Omega(k/\varepsilon^2)$ rows, independent of r.

2.1 Analysis via conditioning

Without loss of generality we henceforth assume $\max\{\|A\|^2, \|A\|_F^2/k\} = \max\{\|B\|^2, \|B\|_F^2/k\} = 1$ (so that $\|A\|^2, \|B\|^2 \le 1$ and $\|A\|_F^2, \|B\|_F^2 \le k$).

We use Lemma 1 in our final analysis to understand the dependence of m on k. Let w, w' each be minimal such that $\|A_{\bar{w}}\|, \|B_{\bar{w'}}\| \le \varepsilon/C'$ for some sufficiently large constant C' (which will be set in the proof of Theorem 1). It was shown that $w, w' = O(k/\varepsilon^2)$ in the proof of Theorem 3.2 (i.b) in [MZ11]. Write the SVDs $A_w = U_{A_w}\Sigma_{A_w}V_{A_w}^T$, $B_{w'} = U_{B_{w'}}\Sigma_{B_{w'}}V_{B_{w'}}^T$.

For $0 \le i \le \log_2(1/\varepsilon^2)$ define D_i' as set of all columns of $U_{A_w}, U_{B_{w'}}$ whose corresponding squared singular values (from $\Sigma_{A_w}, \Sigma_{B_{w'}}$) are at least $1/2^i$. Let D_{A_w} be the set of $\min\{k, w\}$ largest singular vectors from U_{A_w}, and define $D_{B_{w'}}$ similarly. Define $D_i = D_i' \cup D_{A_w} \cup D_{B_{w'}}$. Let s_i denote the dimension of span(D_i), and note the s_i are non-decreasing.

Let $\tilde{s}_i$ be s_i after rounding up to the nearest power of 2. Group all i with the same value of $\tilde{s}_i$ into groups $G_1, G_2, \ldots, G_{\log_2(1/\varepsilon^2)}$. For example if for $i = 0, 1, 2, 3$ the s_i are $3, 4, 15, 16$ then the $\tilde{s}_i$ are $4, 4, 16, 16$ and $G_1 = \{0, 1\}$, $G_2 = \{2, 3\}$. Let v_j be the common value of $\tilde{s}_i$ for i in G_j.

Lemma 2. $\sum_i s_i/2^i \le 8k$.

Proof. Define $s = |D_{A_w} \cup D_{B_{w'}}| \le 2k$ and let s'_i denote the dimension of $\mathrm{span}(D'_i)$. Then the above summation is at most $\sum_i (s/2^i + s'_i/2^i) \le 4k + \sum_i s'_i/2^i$. It thus suffices to bound the second summand by $4k$.

Note that we can find a basis for D'_i among the columns of $U_{A_w}, U_{B_{w'}}$ with corresponding squared singular value at least $1/2^i$, so let $a_i + b_i = s'_i$, where a_i is the number of columns of U_{A_w} in the basis and b_i the number of columns of $U_{B_{w'}}$ in the basis. Then by averaging, if the inequality of the lemma statement does not hold then either $\sum_i a_i/2^i > 2k$ or $\sum_i b_i/2^i > 2k$. Without loss of generality assume the former.

Consider an arbitrary column of U_{A_w}, and suppose it has squared singular value in the range $[1/2^i, 1/2^{i-1})$. Then it is in $\mathrm{span}(D'_j)$ for all $j \ge i$. Its contribution to $\sum_i a_i/2^i$ is therefore $1/2^i + 1/2^{i+1} + \ldots$ which is at most $2/2^i = 1/2^{i-1}$. It follows that $\sum_i a_i/2^i \le 2k$, since the squared Frobenius norm of A_w is at most k. This is a contradiction to $\sum_i a_i/2^i > 2k$. ■

Now we prove the main theorem of this subsection.

Theorem 1. *Suppose that the following conditions hold:*

(1) If $w + w' \le k$, then Π is an ε/C-subspace embedding for the subspace spanned by the columns of $A_w, B_{w'}$. Otherwise if $w + w' > k$, then for each $0 \le i \le \log_2(1/\varepsilon^2)$, Π is an ε_i/C-subspace embedding for $\mathrm{span}(D_{i'})$ with

$$\varepsilon_i = \min\left\{\frac{1}{2}, \varepsilon\sqrt{\frac{v_j}{k}}\right\}$$

where i' is the largest i with s_i in G_j.

(2) $\|\Pi A_{\bar{w}}\|, \|\Pi B_{\bar{w}'}\| \le \varepsilon/C$.

Then Eq. (4) holds as long as C is smaller than some fixed universal constant.

Proof. We would like to bound

$$\|(\Pi A)^T(\Pi B) - A^T B\| \le \underbrace{\|(\Pi A_w)^T \Pi B_{w'} - A_w^T B_{w'}\|}_{\alpha} + \underbrace{\|(\Pi A_{\bar{w}})^T \Pi B_{w'}\|}_{\beta} + \underbrace{\|(\Pi A_w)^T \Pi B_{\bar{w}'}\|}_{\gamma}$$
$$+ \underbrace{\|(\Pi A_{\bar{w}})^T \Pi B_{\bar{w}'}\|}_{\Delta} + \underbrace{\|A_{\bar{w}}^T B_{w'}\|}_{\zeta} + \underbrace{\|A_w^T B_{\bar{w}'}\|}_{\eta} + \underbrace{\|A_{\bar{w}}^T B_{\bar{w}'}\|}_{\Theta} \quad (6)$$

Using $\|XY\| \le \|X\| \cdot \|Y\|$ for any conforming matrices X, Y, we see $\Delta \le \varepsilon^2/C^2$ by condition (2). Furthermore by the definition of w, w' we know $\|A_{\bar{w}}\|, \|B_{\bar{w}'}\| \le \varepsilon/C'$, and thus $\zeta + \eta + \Theta \le 2\varepsilon/C' + (\varepsilon/C')^2$. Note condition (1) implies that Π is a $(1/2)$-subspace embedding for the subspace spanned by columns of $A_w, B_{w'}$ (by taking i maximal). Thus by both conditions we have $\beta, \gamma \le (\varepsilon/C)(1 + 1/2)$.

It only remains to bound α. If $w + w' \le k$, then we are done by condition (1) and Lemma 1. Thus assume $w + w' > k$. Then we have

$$\|(\Pi A_w)^T \Pi B_{w'} - A_w^T B_{w'}\| = \sup_{\|x\|=\|y\|=1} \left| \langle \Pi U_{A_w}\Sigma_{A_w} x, \Pi U_{B_{w'}}\Sigma_{B_{w'}} y\rangle - \langle U_{A_w}\Sigma_{A_w} x, U_{B_{w'}}\Sigma_{B_{w'}} y\rangle \right|$$

Let x, y be any unit norm vectors. Write $x = x^1 + x^2 + \ldots + x^b$ for $b = \log_2(1/\varepsilon^2)$, where x^i is the restriction of x to coordinates for which the corresponding squared singular values in Σ_{A_w} are in $(1/2^i, 1/2^{i-1}]$. Similarly define $y^1, \ldots, y^b$. Then $|\langle \Pi U_{A_w}\Sigma_{A_w}x, \Pi U_{B_{w'}}\Sigma_{B_{w'}}y\rangle - \langle U_{A_w}\Sigma_{A_w}x, U_{B_{w'}}\Sigma_{B_{w'}}y\rangle|$ equals

$$\left|\sum_{i=1}^{b}\sum_{j=1}^{b}\langle \Pi U_{A_w}\Sigma_{A_w}x^i, \Pi U_{B_{w'}}\Sigma_{B_{w'}}y^j\rangle - \langle U_{A_w}\Sigma_{A_w}x^i, U_{B_{w'}}\Sigma_{B_{w'}}y^j\rangle\right|$$
$$\leq \sum_{i=1}^{b}\left|\left\langle \Pi U_{A_w}\Sigma_{A_w}x^i, \Pi U_{B_{w'}}\Sigma_{B_{w'}}\sum_{j\leq i}y^j\right\rangle - \left\langle U_{A_w}\Sigma_{A_w}x^i, \sum_{j\leq i}U_{B_{w'}}\Sigma_{B_{w'}}y^j\right\rangle\right|$$
$$+\sum_{j=1}^{b}\left|\left\langle \Pi U_{A_w}\Sigma_{A_w}\sum_{i\leq j}x^i, \Pi U_{B_{w'}}\Sigma_{B_{w'}}y^j\right\rangle - \left\langle \sum_{i\leq j}x^i, y^j\right\rangle\right| \quad (7)$$

We bound the first sum, as bounding the second is similar. Note $U_{A_w}\Sigma_{A_w}x^i, U_{B_{w'}}\Sigma_{B_{w'}}\sum_{j\leq i}y^j \in D_i$. Therefore by property (1) and Lemma 1,

$$\left|\left\langle \Pi U_{A_w}\Sigma_{A_w}x^i, \Pi U_{B_{w'}}\Sigma_{B_{w'}}\sum_{j\leq i}y^j\right\rangle - \left\langle U_{A_w}\Sigma_{A_w}x^i, U_{B_{w'}}\Sigma_{B_{w'}}\sum_{j\leq i}y^j\right\rangle\right| \leq \frac{\varepsilon_i}{C2^{(i-1)/2}}\cdot\|x^i\|\cdot\|y\|$$
$$\leq \frac{\varepsilon}{C2^{(i-1)/2}}\cdot\sqrt{\frac{2s_i}{k}}\cdot\|x^i\| \quad (8)$$

where Eq. (8) used that the corresponding v value in property (1) is at most $2s_i$. Returning to Eq. (7) and applying Cauchy-Schwarz and Lemma 2,

$$\sum_{i=1}^{b}\left|\left\langle \Pi U_{A_w}\Sigma_{A_w}x^i, \Pi U_{B_{w'}}\Sigma_{B_{w'}}\sum_{j\leq i}y^j\right\rangle - \left\langle U_{A_w}\Sigma_{A_w}x^i, \sum_{j\leq i}U_{B_{w'}}\Sigma_{B_{w'}}y^j\right\rangle\right| \leq \sum_{i=1}^{b}\frac{\varepsilon}{C2^{(i-1)/2}}\cdot\sqrt{\frac{2s_i}{k}}\cdot\|x^i\|$$
$$\leq \frac{2\varepsilon}{C\sqrt{k}}\cdot\left(\sum_{i=1}^{b}\frac{s_i}{2^i}\right)^{1/2}\cdot\left(\sum_{i=1}^{b}\|x^i\|^2\right)^{1/2}$$
$$\leq \frac{2\sqrt{8}\varepsilon}{C}$$

We thus finally have that Eq. (6) is at most $(2\sqrt{8}+3)\varepsilon/C + +(\varepsilon/C)^2 + 2\varepsilon/C' + (\varepsilon/C')^2$, which is at most ε for C, C' sufficiently large constants. ■

Applying Theorem 1:

Example 1: Let Π have $O(k/\varepsilon^2)$ rows forming an orthonormal basis for the span of the columns of $A_w, B_{w'}$. Property (1) is satisfied for every i in fact with $\varepsilon_i = 0$. Property (2) is also satisfied since $\|\Pi A_{\bar{w}}\| \leq \|\Pi\|\cdot\|A_{\bar{w}}\| \leq \varepsilon$, and similarly for bounding $\|\Pi B_{\bar{w'}}\|$.

Example 2: Let Π be $1/\sqrt{m}$ times a random $m\times n$ matrix with independent entries that are subgaussian with variance 1. For example, the entries of Π may be $\mathcal{N}(0, 1/m)$, or uniform in $\{-1/\sqrt{m}, 1/\sqrt{m}\}$. Let m be $\Theta((k+\log(1/\delta))/\varepsilon^2)$. By standard results (see e.g. [CW13]), it is known that such a matrix is an ε-subspace embedding for a k-dimensional subspace with failure probability δ. For property (1) of Theorem 1, if $w + w' \le k$ then we would like Π to be an ε-subspace embedding for a subspace of dimension at most k, which holds with failure probability δ. If $w + w' > k$ then we would like Π to be an ε_i-subspace embedding for $\mathrm{span}(D_{i'})$ for all $1 \le i \le \log_2(1/\varepsilon^2)$ simultaneously. Note $\max_j v_j \le 2(w + w') = O(k/\varepsilon^2)$, and thus $\max_j v_j \le m$. Thus for a subspace under consideration $\mathrm{D}_{\mathrm{i}'}$ for $i' \in G_j$, we have failure probability $\delta^{v_j/k}$ for our choice of m. By construction every v_j is at least k, and the v_j increase at least geometrically. Thus our total failure probability is, by a union bound, $\sum_j \delta^{v_j/k} \le \sum_j \delta^{2^{j-1}} = O(\delta)$. Property (2) of Theorem 1 is satisfied with failure probability δ by [RV13, Theorem 3.2].

Our alternative stable rank AMM analysis in Section 2.2 easily directly applies to the SRHT and sparse subspace embeddings, so we defer our implications for these constructions to that section.

2.2 Analysis via a moment property

Here we provide another way to obtain Eq. (4) for any Π whose subspace embedding property has been established using the moment method, e.g. sparse subspace embeddings [MM13, NN13], dense subgaussian matrices as analyzed in Section A.1, or even the SRHT as analyzed in Section A.2. Our approach in this subsection is inspired by the introduction of the "JL-moment property" in [KN14] to analyze approximate matrix multiplication with Frobenius error. The following is a generalization of [KN14, Definition 6.1], which was only concerned with $d = 1$.

Definition 3. *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,*

$$\mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}} \left\|(\Pi U)^T(\Pi U) - I\right\|^\ell < \varepsilon^\ell \cdot \delta$$

Note that this is just a special case of bounding the expectation of an arbitrary function of $\|(\Pi U)^T(\Pi U) - I\|$. The arguments below will actually apply to any nonnegative, convex, increasing function of $\|(\Pi U)^T(\Pi U) - I\|^2$, but we restrict to moments for simplicity of presentation. The acronym "OSE" refers to *oblivious subspace embedding*, a term coined in [NN13] 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.

Lemma 3. *Suppose $\mathcal{D}$ satisfies the $(\varepsilon, \delta, 2d, \ell)$-OSE moment property and A, B are matrices with (1) the same number of rows, and (2) sum of ranks at most $2d$. Then*

$$\mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}} \left\|(\Pi A)^T(\Pi B) - A^T B\right\|^\ell < \varepsilon^\ell \|A\|^\ell \|B\|^\ell \cdot \delta$$

Proof. First, we apply Lemma 1 to A and B, where U forms an orthonormal basis for the subspace $\mathrm{span}\{\mathrm{columns}(A), \mathrm{columns}(B)\}$, showing that

$$\left\|(\Pi A)^T(\Pi B) - A^T B\right\| \le \left\|(\Pi U)^T(\Pi U) - I\right\| \|A\|\|B\|.$$

Therefore

$$\mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}} \left\|(\Pi A)^T(\Pi B) - A^T B\right\|^\ell \le \mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}} \left\|(\Pi U)^T(\Pi U) - I\right\|^\ell \|A\|^\ell\|B\|^\ell < \varepsilon^\ell \|A\|^\ell\|B\|^\ell \cdot \delta$$

■

Then, just as [KN14, 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. Then, as we will see below, this straightforwardly implies that many OSE constructions can be used in this context, with their number of rows depending on stable rank and not rank.

Theorem 2. *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,*

$$\underset{\Pi \sim \mathcal{D}}{\mathbb{P}}\left(\|(\Pi A)^T(\Pi B) - A^T B\| > \varepsilon\sqrt{(\|A\|^2 + \|A\|_F^2/k)(\|B\|^2 + \|B\|_F^2/k)}\right) < \delta \tag{9}$$

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,

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

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, which does not affect the claim). 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, $\|C\| = \sup_{\|x\|=\|y\|=1} x^T Cy$. 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 $x^T Cy \leq x'^T C' y'$, 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 $\ell \geq 1$),

$$\begin{aligned}
\|\|C'\|_F^2\|_{\ell/2} &= \left\|\sum_{i,j} \|(\Pi A'_i)^T(\Pi B'_j) - A'^T_i B'_j\|^2\right\|_{\ell/2} \\
&\leq \sum_{i,j} \|\|(\Pi A'_i)^T(\Pi B'_j) - A'^T_i B'_j\|^2\|_{\ell/2} \\
&\leq \sum_{i,j} \varepsilon^2 s_i^2 t_j^2 \cdot \delta^{2/\ell} \text{ by Lemma 3} \\
&= \varepsilon^2 \left(\sum_i s_i^2\right) \cdot \left(\sum_j t_j^2\right) \delta^{2/\ell} \\
&\leq \left(\varepsilon\sqrt{(\|A\|^2 + \|A\|_F^2/k)(\|B\|^2 + \|B\|_F^2/k)}\delta^{1/\ell}\right)^2
\end{aligned}$$

Now, $\mathbb{E}\,\|C'\|_F^\ell = \|\|C'\|_F^2\|_{\ell/2}^{\ell/2}$, implying

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

■

Applying Theorem 2: In Section A.1 we show that if Π has independent subgaussian entries, then it satisfies an OSE moment property and thus the analysis in this subsection applies to show that it suffices for such Π to have $m = O((k + \log(1/\delta))/\varepsilon^2)$ rows to satisfy (k, ε)-AMM with failure probability δ. The analyses in [MM13, NN13] when combined with Theorem 2 imply that for sparse subspace embeddings with $s = 1$ non-zero entry per column, one can achieve (k, ε)-AMM with failure probability δ having $m = O(k/(\varepsilon^2\delta))$, although that guarantee was already implied by [KN14, Theorem 6.2]. The analysis in [NN13] combined with Theorem 2 also allows $m = O(k \log^6(k/\delta)/\varepsilon^2)$ and $s = O(\log^3(k/\delta)/\varepsilon)$, and if Conjecture 14 of that work is positively resolved (a conjecture concerning just the settings required to obtain the OSE property) one could even set $m = O((k + \log(1/\delta))/\varepsilon^2)$, $s = O(\log(k/\delta)/\varepsilon)$. Section A.2 shows the SRHT satisfies an OSE moment property and thus one can set $m = O(\varepsilon^{-2}(k + \log(1/(\varepsilon\delta)))\log(k/\delta))$ for that construction. Interestingly our analysis of the SRHT in Section A.2 seems to be asymptotically tighter than any other analyses in previous work even for the basic subspace embedding property.

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 [YPW15] and k-means clustering [CEM$^+$15].

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. [CW09], Theorem 3.1). This will produce

$$\begin{aligned}\tilde{X} &= ((\Pi A)^T \Pi A)^{-1} (\Pi A)^T \Pi B \\ A\tilde{X} &= A((\Pi A)^T \Pi A)^{-1} (\Pi A)^T \Pi B \\ &= U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi B.\end{aligned}$$

Below we give a lemma on the guarantees of the sketched solution in terms of properties of Π.

Theorem 3. *If* Π

1. *satisfies the* $(k, \sqrt{\varepsilon/8})$*-approximate spectral norm matrix multiplication property for* $U_A, P_{\bar{A}}B$
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

$$\|A\tilde{X} - B\|^2 \le (1+\varepsilon)\|P_A B - B\|^2 + (\varepsilon/k) \cdot \|P_A B - B\|_F^2. \tag{10}$$

Proof. We may write:

$$\begin{aligned}\|A\tilde{X} - B\|_2^2 &= \|U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi B - B\|^2 \\ &= \|U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi (P_A B + P_{\bar{A}} B) - P_A B - P_{\bar{A}} B\|^2 \\ &= \|P_A B + U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi P_{\bar{A}} B - P_A B - P_{\bar{A}} B\|^2 \\ &= \|U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi P_{\bar{A}} B - P_{\bar{A}} B\|^2.\end{aligned}$$

So far, we have shown that the error depends only on $P_{\bar{A}}B$ and not $P_A B$ (with the third line following from the fact that the sketched regression is exact on $P_A B$). Now, in the last line, we can see that the two terms lie in orthogonal column spaces (the first in the span of A, the second orthogonal to it). For matrices X and Y with orthogonal column spans, $\|X+Y\|^2 \le \|X\|^2 + \|Y\|^2$, so this is at most

$$\|U_A((\Pi U_A)^T \Pi U_A)^{-1} (\Pi U_A)^T \Pi P_{\bar{A}} B\|^2 + \|P_{\bar{A}} B\|^2.$$

Spectral submultiplicativity then implies the first term is at most

$$(\|U_A\| \cdot \|((\Pi U_A)^T \Pi U_A)^{-1}\| \cdot \|(\Pi U_A)^T \Pi P_{\bar{A}} B\|)^2.$$

$\|U_A\|$ is 1, since U_A is orthonormal. $((\Pi U_A)^T \Pi U_A)^{-1}$ is at most 2, since Π is a subspace embedding for U_A. Finally, $\|(\Pi U_A)^T \Pi P_{\bar{A}} B\|$ is at most

$$\sqrt{\varepsilon/8} \cdot \sqrt{(\|U_A\|^2 + \|U_A\|_F^2/k)(\|P_{\bar{A}}B\|^2 + \|P_{\bar{A}}B\|_F^2/k)} = \sqrt{(\varepsilon/8) \cdot 2 \cdot (\|P_A B - B\|^2 + \|P_A B - B\|^2/k)}.$$

Multiplying these together, squaring, and adding the remaining $\|P_{\bar{A}}B\|^2$ term gives a bound of

$$(1+\varepsilon)\|P_A B - B\|^2 + (\varepsilon/k) \cdot \|P_A B - B\|_F^2$$

as desired. ■

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 matrix 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. [Sar06]):

1. Let $S = \Pi A$.
2. Let P_S be the orthogonal projection operator to the row space of S. Let $\tilde{A} = AP_S$.
3. Compute a singular value decomposition of $\tilde{A}$, and keep only the top k singular vectors. 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 [CW09, Lemma 4.3].

Let A_k be the exact k-truncated SVD approximation of A (and thus the best rank-k approximation, in the spectral and Frobenius norms), and let U_k be the top k column singular vectors, and $A_{\bar{k}} = A - A_k$ be the tail.

Theorem 4. *If* Π

1. *satisfies the* $(k, \sqrt{\varepsilon/8})$*-approximate spectral norm matrix multiplication property for* $U_k, A_{\bar{k}}$
2. *is a* $(1/2)$*-subspace embedding for the column space of* U_k

then

$$\|A - \tilde{A}_k\|^2 \leq (1+\varepsilon)\|A - A_k\|^2 + (\varepsilon/k)\|A - A_k\|_F^2 \tag{11}$$

Proof. Note that this procedure chooses the best possible (in the spectral norm) rank-k approximation to A subject to the constraint of lying in the row space of S. Thus, the spectral norm error can be no worse than the error of a specific such matrix we exhibit.

We simply choose the matrix obtained by running our generalized regression algorithm from A onto U_k, with Π:

$$U_k((\Pi U_k)^T \Pi U_k)^{-1}(\Pi U_k)^T \Pi A$$

This is rank-k by construction, since it is multiplied by U_k, and it lies in the row space of $S = \Pi A$ since that is the rightmost factor. On the other hand, it is an application of the regression algorithm to A where the optimum output is A_k (since that is the projection of A onto the space of U_k). Plugging this into Eq. (10) gives the desired result. ■

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 so that, as n grows, the probability that $\tilde{f}$ is "close" to f^* increases at some good rate.

The recent work [YPW15] 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 [YPW15] considers the case where f^* comes from a Hilbert space $\mathcal{H}$ of functions f such that f is guaranteed to be square integrable, and the map $x \mapsto f(x)$ is a bounded linear functional. The function $\tilde{f}$ is then defined to be the optimal solution to the *Kernel Ridge Regression (KRR)* problem of computing

$$f^{LS} = \underset{f \in \mathcal{H}}{argmin} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \cdot \|f\|_{\mathcal{H}}^2 \right\} \tag{12}$$

for some parameter λ_n. It is known that any $\mathcal{H}$ as above can be written as the closure of the set of all functions

$$g(\cdot) = \sum_{i=1}^{N} \alpha_i k(\cdot, z_i), \tag{13}$$

over all $\alpha \in \mathbb{R}^N$ and vectors $z_1, \ldots, z_N$ for some positive semidefinite *kernel function* k. Furthermore, the optimal solution to Eq. (12) can be expressed as $f^{LS} = \sum_{i=1}^{n} \alpha_i^{LS} \cdot k(\cdot, x_i)$ for some choice of weight vector α^{LS}, and it is known that $\|f^{LS} - f^*\|_n$ will be small with high probability, over the randomness in w, if λ_n is chosen appropriately (see [YPW15] for background references and precise statements).

After rewriting Eq. (12) using Eq. (13) and defining a matrix K with $K_{i,j} = k(x_i, x_j)$, one arrives at a reformulation for KRR of 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 [YPW15] 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 [YPW15, Section 4.1.2], specifically equation (26) in that work):

- Π is a $(1/2)$-subspace embedding for a particular low-dimensional subspace
- $\|\Pi B\| = O(\|B\|)$ for a particular matrix B of low stable rank (B is UD_2 in [YPW15]). Note

  $$\|\Pi B\| = \|(\Pi B)^T \Pi B\|^{1/2} \le \left(\|(\Pi B)^T \Pi B - B^T B\| + \|B^T B\|\right)^{1/2} \le \|(\Pi B)^T \Pi B - B^T B\|^{1/2} + \|B\|,$$

  and thus it suffices for Π to provide the approximate matrix multiplication property for the product $B^T B$, where B has low stable rank.

The first bullet simply requires a subspace embedding in the standard sense, and for the second bullet [YPW15] avoided AMM by obtaining a bound on $\|\Pi B\|$ directly by their own analyses for gaussian and the SRHT (in the gaussian case, it also follows from [RV13, Theorem 3.2]). 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 [BZMD15, CEM+15], 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 $k \ge 1$, and the goal is to find k points $y_1, \ldots, y_k \in \mathbb{R}^d$ minimizing

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

That is, the n points can be partitioned arbitrarily into k clusters, then a "cluster center" should be assigned to each cluster so as to minimize sums of squared Euclidean distances of each of the n points to their cluster centers. It is a standard fact that once a partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of the n points into clusters is fixed, the optimal cluster centers to choose are the centroids of the points in each of the k partitions, i.e. $y_j = (1/|P_j|) \cdot \sum_{i \in P_j} x_i$.

One key observation common to both of the works [BZMD15, CEM+15] is that k-means clustering is 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

$$(X_{\mathcal{P}})_{i,j} = \begin{cases} \frac{1}{\sqrt{|P_j|}}, & \text{if } i \in P_j \\ 0, & \text{otherwise} \end{cases}$$

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

$$\mathrm{argmin}_{\mathcal{P}} \|A - X_{\mathcal{P}} X_{\mathcal{P}}^T A\|_F^2$$

where $\mathcal{P}$ ranges over all partitions of $\{1, \ldots, n\}$ into k sets. It is easy to verify that the non-zero 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 at most 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

$$\mathrm{argmin}_{P \in \mathcal{S}} \|(I - P)A\|_F^2. \tag{14}$$

One can verify this by hand, since the rows of A are the points x_i, and the ith row of PA for $P = X_{\mathcal{P}} X_{\mathcal{P}}^T$ is the centroid of the points in i's partition in $\mathcal{P}$.

The work [CEM$^+$15] 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 $\tilde{P} \in \mathcal{S}$ satisfies

$$\|(I - \tilde{P})A\Pi^T\|_F^2 \le \gamma \cdot \min_{P\in\mathcal{S}} \|(I - P)A\Pi^T\|_F^2, \tag{15}$$

then

$$\|(I - \tilde{P})A\|_F^2 \le \frac{(1+\varepsilon)}{(1-\varepsilon)} \cdot \gamma \cdot \min_{P\in\mathcal{S}} \|(I - P)A\|_F^2. \tag{16}$$

One set of sufficient conditions for Π is as follows (see [CEM$^+$15, Lemma 10]). Let A_k denote the best rank-k approximation to A and let $A_{\bar{k}} = A - A_k$. Define $Z \in \mathbb{R}^{d\times r}$ for $r = 2k$ by $Z = V_r$, i.e. the top r right singular vectors of A are the columns of Z. Define $B_1 = Z^T$ and $B_2 = \frac{\sqrt{k}}{\|A_{\bar{k}}\|_F} \cdot (A - AZZ^T)$. Define $B \in \mathbb{R}^{(n+r)\times d}$ as having B_1 as its first r rows and B_2 as its lower n rows. Then [CEM$^+$15, Lemma 10] states that Eq. (15) implies Eq. (16) as long as

$$\|(\Pi B^T)^T(\Pi B^T) - BB^T\| < \varepsilon, \tag{17}$$

$$\text{and } \left|\|\Pi B_2\|_F^2 - \|B_2\|_F^2\right| \le \varepsilon k \tag{18}$$

One can easily check $\|B\|^2 = 1$ and $\|B\|_F^2 \le 3k$, so the stable rank $\tilde{r}(B)$ is at most $3k$. Thus Eq. (17) is implied by the $(3k, \varepsilon/2)$-AMM property for B^T, B^T, and our results apply to show that Π can be taken to have $m = O((k + \log(1/\delta))/\varepsilon^2)$ rows to have success probability $1-\delta$ for Eq. (17). Obtaining Eq. (18) is much simpler and can be derived from the JL moment property (see the proof of [KN14, Theorem 6.2]).

Without our results on stable-rank AMM provided in this current work, [CEM$^+$15] gave a different analysis, avoiding [CEM$^+$15, Lemma 10], which required Π to have $m = \Theta(k \cdot \log(1/\delta)/\varepsilon^2)$ 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, approximate matrix multiplication (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. This is analyzed using matrix Chernoff bounds [AW02], and for the case of k-dimensional subspace embedding or approximate matrix multiplication of rank-k matrices, it can produce $O(k(\log k)/\varepsilon^2)$ samples [SS11]. The second method is the deterministic selection method given in [BSS12], often called "BSS", choosing only $O(k/\varepsilon^2)$ rows. This still runs in polynomial time, but requires many relatively expensive linear algebra steps and thus is slower in general.

The matrix Chernoff methods can be extended to the stable-rank case, making even the log factor depend only on the stable rank, using "intrinsic dimension" variants of the bounds as presented in Chapter 7 of [Tro15]. Specifically, Theorem 6.3.1 of that work can be applied with each

n summands each equal to $\frac{1}{n}\left(\frac{1}{p_i}a_i^T b_i - A^T B\right)$, where a_i is the ith row of A, and i is random with the probability of choosing a particular row i equal to

$$p_i = \frac{\|a_i\|^2 + \|b_i\|^2}{\sum_j \|a_j\|^2 + \|b_j\|^2}$$

We here give an extension of BSS that covers low stable rank matrices as well.

Theorem 5. *Given an n by d matrix A such that $\|A\|^2 \leq 1$ and $\|A\|_F^2 \leq k$, and an $\varepsilon \in (0,1)$, there exists a diagonal matrix S with $O(k/\varepsilon^2)$ nonzero entries such that*

$$\|(SA)^T(SA) - A^T A\| \leq \varepsilon$$

Such an S can be computed by a polynomial-time algorithm.

When $A^T A$ is the identity, this is just the original BSS result. It is also stronger than Theorem 3.3 of [KMST10], 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 5, but with the row count scaling as k/ε^3 rather than k/ε^2.

Proof. The proof closely follows the original proof of BSS. However, for simplicity, and because the tight constants are not needed for most applications, we do not include [BSS12, Claim 3.6] and careful parameter-setting.

At each step, the algorithm will maintain a partial approximation $Z = (SA)^T(SA)$ (the matrix "A" in [BSS12]), with S beginning as 0. Additionally, we keep track of upper and lower "walls" X_u and X_l; in the original BSS these are just multiples of the identity. The final S will be returned by the algorithm (rescaled by a constant so that the average of the upper and lower walls is $A^T A$).

We will maintain the invariants

$$tr(A(X_u - Z)^{-1}A^T) \leq 1 \tag{19}$$

$$tr(A(Z - X_l)^{-1}A^T) \leq 1. \tag{20}$$

These are the so-called upper and lower potentials from BSS. We also require $X_u \prec Z \prec X_l$; recall $M \prec M'$ means that $M' - M$ is positive definite. Note that unlike [BSS12], here we do not apply a change of variables making $A^T A$ the identity (to avoid confusion, since that would change the Frobenius norm). This is the reason for the slightly more complicated form of the potentials.

In the original BSS, X_u and X_l were always scalar multiples of the identity (here, without the change of variables, that would correspond to always being multiples of $A^T A$). [BSS12] thus simply represented them with scalars. Like BSS, we will increase X_u and X_l by multiples of $A^T A$–however, the key difference from BSS is that they are *initialized* to multiples of the identity, rather than $A^T A$. In particular, we may initialize X_u to kI and X_l to $-kI$. This is still good enough to get the spectral norm bounds we require here (as opposed to the stronger multiplicative approximation guaranteed by BSS).

We will have two scalar values, δ_u and δ_l, depending only on ε; they will be set later. One step consists of

1. Choose a row a_i from A and a positive scalar t, and add $ta_i a_i^T$ to Z (via increasing the i component of S).

2. Add $\delta_u A^T A$ to X_u and $\delta_l A^T A$ to X_l.

We will show that with suitable values of δ_u and δ_l, for any Z obeying the invariants there always exists a choice of i and t such that the invariants will still be true after the step is complete. This corresponds to Lemmas 3.3 through 3.5 of BSS.

For convenience, we define, at a given step, the matrix functions of y

$$\begin{aligned} M_u(y) &= ((X_u + yA^T A) - Z)^{-1} \\ M_l(y) &= (Z - (X_l + yA^T A))^{-1}. \end{aligned}$$

The upper barrier value, after making a step of $ta_i a_i^T$ and increasing X_u, is

$$tr(A((X_u + \delta_u A^T A) - (Z + ta_i a_i^T))^{-1} A^T).$$

Applying the Sherman-Morrison formula, and cyclicity of trace, to the rank-1 update $ta_i a_i^T$, this can be rewritten as

$$tr(AM_u(\delta_u)A^T) + \frac{ta_i^T M_u(\delta_u) A^T A M_u(\delta_u) a_i}{1 - ta_i^T M_u(\delta_u) a_i}.$$

Since the function $f(y) = tr(AM_u(y)A^T)$ is a convex function of y with derivative

$$f'(y) = -\, tr(AM_u(y)A^T AM_u(y)A^T),$$

we have $f(\delta_u) - f(0) \leq -\delta_u \, tr(AM_u(\delta_u)A^T AM_u(\delta_u)A^T)$. Then the difference between the barrier before and after the step is at most

$$\frac{ta_i^T M_u(\delta_u) A^T A M_u(\delta_u) a_i}{1 - ta_i^T M_u(\delta_u) a_i} - \delta_u \, tr(AM_u(\delta_u)A^T AM_u(\delta_u)A^T).$$

Constraining this to be no greater than zero, rewriting in terms of $\frac{1}{t}$ and pulling it out gives

$$\frac{1}{t} \geq \frac{a_i^T M_u(\delta_u) A^T A M_u(\delta_u) a_i}{\delta_u \, tr(AM_u(\delta_u)A^T AM_u(\delta_u)A^T)} + a_i^T M_u(\delta_u) a_i.$$

Furthermore, as long as $\frac{1}{t}$ is at least this, Z will remain below X_u, since the barrier must approach infinity as t approaches the smallest value passing X_u.

For the lower barrier value after the step, we get

$$tr(A((Z + ta_i a_i^T) - (X_l + \delta_l A^T A))^{-1} A^T).$$

Again, applying Sherman-Morrison rewrites it as

$$tr(AM_l(\delta_l)A^T) - \frac{ta_i^T M_l(\delta_l) A^T A M_l(\delta_l) a_i}{1 + ta_i^T M_l(\delta_l) a_i}.$$

Again, due to convexity the increase in the barrier from raising X_l is at most δ_l times the local derivative. The difference in the barrier after the step is then at most

$$-\frac{ta_i^T M_l(\delta_l) A^T A M_l(\delta_l) a_i}{1 + tM_l(\delta_l) a_i} + \delta_l \, tr(AM_l(\delta_l)A^T AM_l(\delta_l)A^T).$$

This is not greater than zero as long as

$$\frac{1}{t} \leq \frac{a_i^T M_l(\delta_l) A^T A M_l(\delta_l) a_i}{\delta_l \, tr(A M_l(\delta_l) A^T A M_l(\delta_l) A^T)} - a_i^T M_l(\delta_l) a_i.$$

There is some value of t that works for a_i as long as the lower bound for $\frac{1}{t}$ is no larger than the upper bound. To show that there is at least one choice of i for which this holds, we look at the sum of all the lower bounds and compare to the sum of all the upper bounds. Summing the former over all i gets

$$\frac{tr(A M_u(\delta_u) A^T A M_u(\delta_u) A^T)}{\delta_u \, tr(A M_u(\delta_u) A^T A M_u(\delta_u) A^T)} + tr(A M_u(\delta_u) A^T)$$

and the latter gets

$$\frac{tr(A M_l(\delta_l) A^T A M_l(\delta_l) A^T)}{\delta_l \, tr(A M_l(\delta_l) A^T A M_l(\delta_l) A^T)} - tr(A M_l(\delta_l) A^T).$$

Finally, note that

$$tr(A M_u(\delta_u) A^T) = tr(A((X_u + \delta_u A^T A) - Z)^{-1} A^T) \leq tr(A(X_u - Z)^{-1} A^T) \leq 1$$

and the lower barrier implies $Z - X_l \succ A^T A$, implying that as long as $\delta_l \leq \frac{1}{2}$,

$$tr(A M_l(\delta_l) A^T) = tr(A(Z - (X_l + \delta_l A^T A))^{-1} A^T) \leq 2 \, tr(A(Z - X_l)^{-1} A^T) \leq 2.$$

Thus, we can always make a step as long as δ_u and δ_l are set so that

$$\frac{1}{\delta_u} + 1 \leq \frac{1}{\delta_l} - 2$$

and $\delta_l \leq \frac{1}{2}$. This is satisfied by

$$\begin{aligned} \delta_u &= \varepsilon + 2\varepsilon^2 \\ \delta_l &= \varepsilon - 2\varepsilon^2. \end{aligned}$$

Before the first step, X_u and X_l can be initialized as kI and $-kI$, respectively. If the algorithm is then run for $\frac{k}{\varepsilon^2}$ steps, we have:

$$\begin{aligned} X_u &= \frac{k}{\varepsilon} A^T A + 2k A^T A + kI \\ &\preceq \frac{k}{\varepsilon} A^T A + 3kI \\ X_l &= \frac{k}{\varepsilon} A^T A - 2k A^T A - kI \\ &\succeq \frac{k}{\varepsilon} A^T A - 3kI. \end{aligned}$$

$\frac{\varepsilon}{k} X_u$ and $\frac{\varepsilon}{k} X_l$ both end up within $3\varepsilon I$ of $A^T A$, so $\frac{\varepsilon}{k} Z$ (from $\sqrt{\frac{\varepsilon}{k}} S$) satisfies the requirements of the output for 3ε (one can simply apply this argument for $\varepsilon/3$). Furthermore, all the computations required to verify the preservation of invariants and compute explicit ts can be performed in polynomial time. ■

This obtains more general AMM as a corollary:

Corollary 1. *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.*

Proof. Apply Theorem 5 to a matrix X consisting of the columns of $\frac{A}{\sqrt{2}\max(\|A\|_2, \|A\|_F/\sqrt{k})}$ appended to the columns of $\frac{B}{\sqrt{2}\max(\|B\|_2, \|B\|_F/\sqrt{k})}$, and use the resulting S.

Note that X satisfies the conditions of that theorem, since concatenating the sets of columns at most adds the squares of their spectral and Frobenius norms. $(SA)^T(SB) - A^T B$ is a submatrix of $2\max(\|A\|_2, \|A\|_F/\sqrt{k})\max(\|B\|_2, \|B\|_F/\sqrt{k})((SX)^T(SX) - X^T X)$, so its spectral norm is upper bounded by the spectral norm of that matrix, which in turn is bounded by the guarantee of Theorem 5. ■

Acknowledgments

We thank Jarosław Błasiok for pointing out the connection between low stable rank approximate matrix multiplication and the analyses in [YPW15].

References

[AC09] Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[AL13] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013.

[AW02] Rudolf Ahlswede and Andreas J. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[BSS12] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

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

[CEM+15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina 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`.

[CLL+10] 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.

[CLM+15] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 6th Annual Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 181–190, 2015.

[CW09] 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. Full version at `http://researcher.watson.ibm.com/researcher/files/us-dpwoodru/cw09.pdf`.

[CW13] 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`.

[DKM06] 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.

[DKS10] 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, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 341–350, 2010.

[DMMW12] 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.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[HMT11] 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.

[HW71] 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.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[KMN11] Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit johnson-lindenstrauss families. In *Proceedings of the 14th International Workshop on Randomization and Computation (RANDOM)*, pages 628–639, 2011.

[KMST10] Alexandra Kolla, Yury Makarychev, Amin Saberi, and Shang-Hua Teng. Subgraph sparsification and nearly optimal ultrasparsifiers. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 57–66, New York, NY, USA, 2010. ACM.

[KN14] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4, 2014.

[KVZ14] Anastasios T. Kyrillidis, Michail Vlachos, and Anastasios Zouzias. Approximate matrix multiplication with application to linear embeddings. *CoRR*, abs/1403.7683, 2014.

[KW11] 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.

[LBKW14] 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)*, pages 3113–3121, 2014.

[LDFU13] 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.

[LMP13] 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)*, pages 127–136, 2013.

[LP86] François Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

[LPP91] François Lust-Piquard and Gilles Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[MM13] 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.

[MZ11] 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.

[NN13] Jelani Nelson and Huy L. Nguyen. 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.

[NN14] Jelani Nelson and Huy L. Nguyên. Lower bounds for oblivious subspace embeddings. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 883–894, 2014.

[NPW14] 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.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[Sar06] 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.

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

[Tro11] Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal.*, 3(1–2):115–126, 2011.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[TZ12] 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.

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

[YPW15] Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *CoRR*, abs/1501.06195, 2015.

Appendix

A OSE moment property

In the following two subsections we show the OSE moment property for both subgaussian matrices and the SRHT.

A.1 Subgaussian matrices

In this section, we show the OSE moment property for distributions satisfying a JL condition, namely the JL moment property. This includes matrices with i.i.d. entries that are mean zero and subgaussian with variance $1/m$.

Definition 4. *[KMN11]* Let $\mathcal{D}$ be a distribution over $\mathbb{R}^{m\times n}$. We say $\mathcal{D}$ has the (ε, δ, p)-*JL moment property* if for all $x \in \mathbb{R}^n$ of unit norm,

$$\mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}} |\|\Pi x\|^2 - 1|^p < \varepsilon^p \cdot \delta.$$

The following theorem follows from the proof of Lemma 8 in the full version of [CW13]. We give a different proof here inspired by the proof of [FR13, Theorem 9.9], which is slightly shorter and more self-contained. A weaker version appears in [Sar06, Lemma 10], where the size bound on X is $(Cd/\varepsilon)^d$ for a constant $C \geq 1$ instead of simply C^d.

Theorem 6. *Let $U \in \mathbb{R}^{n\times d}$ with orthonormal columns be arbitrary. Then there exists a set $X \subset \mathbb{R}^n$, $|X| \leq 9^d$, each of norm at most 1 such that*

$$\|(\Pi U)^T(\Pi U) - I\| \leq 2 \cdot \sup_{x\in X} |\|\Pi x\|^2 - 1|$$

Proof. We will show that if $\sup_{x\in X} |\|\Pi x\|^2 - 1| < \varepsilon/2$ then $\|(\Pi U)^T(\Pi U) - I\| < \varepsilon$, where $\varepsilon > 0$ is some positive real. Define $A = (\Pi U)^T(\Pi U) - I$. Since A is symmetric,

$$\|A\| = \sup_{\|x\|=1} |x^T Ax| = \sup_{\|x\|=1} | \langle Ax, x\rangle |$$

Let T_γ be a finite γ-net of ℓ_2^d, i.e. $T_\gamma \subset \ell_2^d$ and for every $x \in \mathbb{R}^d$ of unit norm there exists a $y \in T_\gamma$ such that $\|x - y\|_2 \le \gamma$. As we will see soon, there exists such a T_γ of size at most $(1 + 2/\gamma)^d$. We will show that if Π satisfies the JL condition on $T' = \{Uy : y \in T_{1/4}\}$ with error $\varepsilon/2$, then $\|A\| < \varepsilon$; that is, $(1 - \varepsilon/2)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1 + \varepsilon/2)\|x\|_2^2$ for all $x \in T'$.

Let x be a unit norm vector that achieves the sup above, i.e. $\|A\| = | \langle Ax, x\rangle |$. Then, letting y be the closest element of T_γ to x,

$$\begin{aligned}\|A\| &= | \langle Ax, x\rangle | \\ &= | \langle Ay, y\rangle + \langle A(x + y), x - y\rangle | \\ &\le \frac{\varepsilon}{2} + \|A\| \cdot \|x + y\| \cdot \|x - y\| \\ &\le \frac{\varepsilon}{2} + 2\gamma\|A\|.\end{aligned}$$

Rearranging gives $\|A\| \le \varepsilon/(2(1 - 2\gamma))$, which is ε for $\gamma = 1/4$.

Now we must show that we can take $|T_\gamma| \le (1 + 2/\gamma)^d$. The following is a standard covering/packing argument for bounding metric entropy. Imagine packing as many radius-$(\gamma/2)$ ℓ_2 balls as possible into $\mathbb{R}^d$, centered at points with at most unit norm and such that these balls do not intersect each other. Then these balls all fit into a radius-$(1 + \gamma/2)$ ℓ_2 ball centered at the origin, and thus the number of balls we have packed is at most the ratio of the volume of a $(1 + \gamma/2)$ ball to the volume of a $\gamma/2$ ball, which is $((1 + \gamma/2)/(\gamma/2))^d = (1 + 2/\gamma)^d$. Now, take those maximally packed radius-$(\gamma/2)$ balls and double each of their radii to be radius γ. Then every point in the unit ball is contained in at least one of these balls by the triangle inequality, which is exactly the property we wanted from T_γ (T_γ is just the centers of these balls). To see why every point is in at least one such ball, if some $x \in \mathbb{R}^d$ of unit norm is not contained in any doubled ball then a $\gamma/2$-ball about x would be disjoint from our maximally packed $\gamma/2$ balls, a contradiction. ∎

Lemma 4. *If $\mathcal{D}$ satisfies the (ε, δ, p)-JL moment property, then $\mathcal{D}$ satisfies the $(2\varepsilon, 9^d\delta, d, p)$-OSE moment property*

Proof. By Theorem 6, there exists a subset $X \subset \mathbb{R}^n$ of at most 9^d points such that

$$\begin{aligned}\mathbb{E}\,\|(\Pi U)^T(\Pi U) - I\|^p &\le 2^p \cdot \mathbb{E} \sup_{x\in X} |\|\Pi x\|^2 - 1|^p \\ &\le 2^p \cdot \sum_{x\in X} \mathbb{E}\,|\|\Pi x\|^2 - 1|^p \\ &\le 2^p \cdot 9^d \cdot \varepsilon^p \cdot \delta \\ &= (2\varepsilon)^p \cdot 9^d\delta.\end{aligned}$$

∎

It is known that if $\mathcal{D}$ is a distribution over $\mathbb{R}^{m\times n}$ with $m = \Omega(\log(1/\delta)/\varepsilon^2)$ and for $\Pi \sim \mathcal{D}$, the entries of Π are independent subgaussians with mean zero and variance $1/m$, then $\mathcal{D}$ has the $(\varepsilon/2, \delta, \Theta(\log(1/\delta)))$-JL moment property [KMN11]. Thus such a matrix has the $(\varepsilon, \delta, d, \Theta(d + \log(1/\delta)))$-OSE moment property for $\delta < 2^{-d}$ by Lemma 4.

A.2 Subsampled Randomized Hadamard Transform (SRHT)

Recall the SRHT is the $m \times n$ matrix $\Pi = (1/\sqrt{m}) \cdot SHD$ for n a power of 2 where D has diagonal entries $\alpha_1, \ldots, \alpha_n$ that are independent and uniform in $\{-1, 1\}$, H is the unnormalized Hadamard transform with $H_{i,j} = (-1)^{\langle i,j \rangle}$ (treating i, j as elements of the vector space $\mathbb{F}_2^{\log_2 n}$), and S is a sampling matrix. That is, the rows of S are independent, and each row has a 1 in a uniformly random location and zeroes elsewhere. A similar construction is where S is an $n \times n$ diagonal matrix with $S_{i,i} = \eta_i$ being independent Bernoulli random variables each of expectation m/n (so that, in expectation, S selects m rows from HD). We will here show the moment property for this latter variant since it makes the notation a tad cleaner, though the analysis we present holds essentially unmodified for the former variant as well.

Our analysis below implies that the SRHT provides an ε-subspace embedding for d-dimensional subspaces with failure probability δ for $m = O(\varepsilon^{-2}(d + \log(1/(\varepsilon\delta))) \log(d/\delta))$. This is an improvement over analyses we have found in previous works. The analysis in [Tro11] only considers constant ε and $\delta = O(1/d)$ and for these settings achieves $m = O((d+\log n) \log d)$, which is still slightly worse than our bound for this setting of ε, δ (our bound removes the $\log n$ and achieves any $1/poly(d)$ failure probability with the same m). The analysis in [LDFU13] only allows failure probabilities greater than n/e^d. They show failure probability $\delta + n/e^d$ is achieved for $m = O(d \log(d/\delta)/\varepsilon^2)$, which is also implied by our result if $m \le n$ (which is certainly the case in applications for the SRHT to be useful, since otherwise one could use the $n \times n$ identity matrix as a subspace embedding). The reason for these differences is that previous works operate by showing HDU has small row norms with high probability over D; since there are n rows, some logarithmic dependence on n shows up in a union bound. After this conditioning, one then shows that S works. Our analysis does not do any such conditioning at all. Interestingly, such a conditioning approach was done even for the case $d = 1$ [AC09]. As we see below, this approach is slightly lossy (essentially the $\log n$ terms that appear from the conditioning approach can be very slightly improved to $\log m$).

Our main motivation in re-analyzing the SRHT was not to improve the bounds, but simply to give an analysis demonstrating that the SRHT satisfies the OSE moment property. The fact that our moment based analysis below (very slightly) improved m was a fortunate accident. Before we present our proof of the OSE moment property for the SRHT, we state a theorem we will use. For a random matrix M, we henceforth use $\|M\|_p$ to denote $(\mathbb{E}\,\|M\|_{S_p}^p)^{1/p}$ where $\|M\|_{S_p}$ is the Schatten-p norm, i.e. the ℓ_p norm of the singular values of M.

Theorem 7 (Non-commutative Khintchine inequality [LP86, LPP91]). *Let $X_1, \ldots, X_n$ be fixed real matrices and $\sigma_1, \ldots, \sigma_n$ be independent Rademachers. Then*

$$\forall p \ge 1,\ \|\sum_i \sigma_i X_i\|_p \lesssim \sqrt{p} \cdot \max\left\{ \|(\sum_i X_i X_i^T)^{1/2}\|_{S_p}, \|(\sum_i X_i^T X_i)^{1/2}\|_{S_p} \right\}.$$

We will also make use of the Hanson-Wright inequality.

Theorem 8 (Hanson-Wright [HW71]). *For (σ_i) independent Rademachers and A symmetric,*

$$\forall p \ge 1,\ \|\sigma^T A \sigma - \mathbb{E}\,\sigma^T A \sigma\|_p \lesssim \sqrt{p} \cdot \|A\|_F + p \cdot \|A\|.$$

Note that for scalar random variables X, it holds that $\|X\|_p \le \|X\|_q$ whenever $p < q$. This is not true for the random matrix norm $\|M\|_p = (\mathbb{E}\,\|M\|_{S_p}^p)^{1/p}$ (as a simple counter-example, consider M being the identity matrix with probability 1). We use the following lemma instead.

Lemma 5. *Let M be a random matrix of rank at most r. Also suppose $1 \le p < q < \infty$. Then*

$$\|M\|_p \le r^{1/p-1/q} \cdot \|M\|_q$$

Proof. Consider the scalar random variable α distributed as follows:

$$\alpha = \begin{cases} \text{uniformly random singular value of } M, & \text{w.p. } rank(M)/r \\ 0, & \text{otherwise} \end{cases}$$

Then $\|\alpha\|_p \le \|\alpha\|_q$, i.e. $(r^{-1}\,\mathbb{E}_M \|M\|_{S_p}^p)^{1/p} \le (r^{-1}\,\mathbb{E}_M \|M\|_{S_q}^q)^{1/q}$. The lemma follows. ■

We now present our main analysis of this subsection.

Theorem 9. *The SRHT satisfies the $(\varepsilon, \delta, d, p)$-moment property for $p = \log(d/\delta)$ as long as $m \gtrsim \varepsilon^{-2}(d\log(d/\delta) + \log(d/\delta)\log(m/\delta)) \simeq \varepsilon^{-2}(d + \log(1/(\varepsilon\delta)))\log(d/\delta)$.*

Proof. For a fixed $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, we would like to bound

$$\mathop{\mathbb{E}}_{\alpha,\eta} \|\frac{1}{m}(SHDU)^T(SHDU) - I\|^p.$$

Since $p \ge \log d$ we have

$$\|\frac{1}{m}(SHDU)^T(SHDU) - I\| \simeq \|\frac{1}{m}(SHDU)^T(SHDU) - I\|_{S_p} \tag{21}$$

by Hölder's inequality. Also, let $z_1, \ldots, z_n$ be the rows of HDU, as column vectors, so that

$$\frac{1}{m}(SHDU)^T(SHDU) = \frac{1}{m}\sum_{i=1}^n \eta_i z_i z_i^T. \tag{22}$$

Note also $\sum_i z_i z_i^T = (HDU)^T HDU = n \cdot I$ for any D, so the identity matrix is the expectation, over η, of the right hand side of Eq. (22) for any D. Thus we are left wanting to bound

$$\|\frac{1}{m}\sum_i \eta_i z_i z_i^T - \mathop{\mathbb{E}}_{\eta'}\frac{1}{m}\sum_i \eta_i' z_i z_i^T\|_p$$

where the η_i' are identically distributed as the η_i but independent of them. Below we use $\|f(X)\|_{L^p(X)}$ to denote $(\mathbb{E}_X |f(X)|^p)^{1/p}$. Thus for (σ_i) independent Rademachers,

$$\begin{aligned}
\|\frac{1}{m}\sum_i \eta_i z_i z_i^T - I\|_p &= \|\frac{1}{m}\sum_i \eta_i z_i z_i^T - \mathop{\mathbb{E}}_{\eta'}\frac{1}{m}\sum_i \eta_i' z_i z_i^T\|_p \qquad (23)\\
&= \|\|\frac{1}{m}\sum_i \eta_i z_i z_i^T - \mathop{\mathbb{E}}_{\eta'}\frac{1}{m}\sum_i \eta_i' z_i z_i^T\|_{L^p(\eta)}\|_{L^p(\alpha)}\\
&\le \frac{1}{m}\|\|\sum_i \eta_i z_i z_i^T - \sum_i \eta_i' z_i z_i^T\|_{L^p(\eta,\eta')}\|_{L^p(\alpha)} \text{ (Jensen's inequality)}\\
&= \frac{1}{m}\cdot\|\sum_i(\eta_i - \eta_i')z_i z_i^T\|_p
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m} \cdot \| \sum_i \sigma_i(\eta_i - \eta_i')z_i z_i^T \|_p \text{ (equal in distribution)}\\
&\leq \frac{2}{m} \cdot \| \sum_i \sigma_i \eta_i z_i z_i^T \|_p \text{ (triangle inequality)}\\
&\lesssim \frac{\sqrt{p}}{m} \cdot \|(\sum_i \eta_i \|z_i\|_2^2 \cdot z_i z_i^T)^{1/2}\|_p \text{ (Theorem 7)}\\
&= \frac{\sqrt{p}}{m} \cdot \| \sum_i \eta_i \|z_i\|_2^2 \cdot z_i z_i^T \|_{p/2}^{1/2}\\
&\leq \sqrt{d^{1/p}} \cdot \frac{\sqrt{p}}{m} \cdot \| \sum_i \eta_i \|z_i\|_2^2 \cdot z_i z_i^T \|_p^{1/2} \text{ (Lemma 5)}\\
&\lesssim \frac{\sqrt{p}}{m} \cdot \|(\max_i \eta_i \|z_i\|_2^2) \cdot (\sum_i \eta_i z_i z_i^T)\|_p^{1/2} \text{ (since } d^{1/p} \leq 2)\\
&= \frac{\sqrt{p}}{m} \cdot \left(\mathop{\mathbb{E}}_{\alpha,\eta} ((\max_i \eta_i \|z_i\|_2^2)^{p/2} \cdot \| \sum_i \eta_i z_i z_i^T \|_{S_p}^{p/2}) \right)^{1/p}\\
&\leq \frac{\sqrt{p}}{m} \cdot \left((\mathop{\mathbb{E}}_{\alpha,\eta} \max_i \eta_i \|z_i\|_2^2)^p)^{1/2} \cdot (\mathop{\mathbb{E}}_{\alpha,\eta} \| \sum_i \eta_i z_i z_i^T \|_{S_p}^p)^{1/2} \right)^{1/p} \quad \text{(Cauchy-Schwarz)}\\
&= \frac{\sqrt{p}}{m} \cdot \| \max_i \eta_i \|z_i\|_2^2 \|_p^{1/2} \cdot \| \sum_i \eta_i z_i z_i^T \|_p^{1/2}\\
&\leq \sqrt{\frac{p}{m}} \cdot \| \max_i \eta_i \|z_i\|_2^2 \|_p^{1/2} \cdot (d^{1/p} + \| \frac{1}{m} \sum_i \eta_i z_i z_i^T - I \|_p^{1/2}) \text{ (triangle inequality)}
\end{aligned} \tag{24}
$$

By choice of $p \geq \log d$, note $1 \leq d^{1/p} \leq 2$. Letting Q denote $\|\frac{1}{m}\sum_i \eta_i z_i z_i^T - I\|_p^{1/2}$ and R denote $\sqrt{p/m} \cdot \| \max_i \eta_i \|z_i\|_2^2 \|_p^{1/2}$, combining Eq. (23) and Eq. (24) we have

$$Q^2 \lesssim R + RQ$$

implying that for some fixed constant $C > 0$, we have $Q^2 - CRQ - CR \leq 0$. This implies that Q is at most the larger root of the associated quadratic equation, i.e. $Q \lesssim \max\{\sqrt{R}, R\}$, or equivalently

$$\|\frac{1}{m} \sum_i \eta_i z_i z_i^T - I\|_p \lesssim \max\{R, R^2\} \tag{25}$$

It only remains to bound R, which in turn amounts to bounding $\| \max_i \eta_i \|z_i\|_2^2 \|_p^{1/2}$. Define $q = \max\{p, \log m\}$, and note $\| \cdot \|_p \leq \| \cdot \|_q$. Then

$$
\begin{aligned}
\| \max_i \eta_i \|z_i\|_2^2 \|_q &= \left(\mathop{\mathbb{E}}_{\alpha,\eta} \max_i \eta_i^q (\|z_i\|_2^2)^q \right)^{1/q}\\
&\leq \left(\mathop{\mathbb{E}}_{\alpha,\eta} \sum_i \eta_i^q (\|z_i\|_2^2)^q \right)^{1/q}
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\sum_i \mathop{\mathbb{E}}_{\alpha,\eta} \eta_i^q (\|z_i\|_2^2)^q \right)^{1/q} \\
&\le \left(n \cdot \max_i \mathop{\mathbb{E}}_{\alpha,\eta} \eta_i^q (\|z_i\|_2^2)^q \right)^{1/q} \\
&= \left(n \cdot \max_i (\mathop{\mathbb{E}}_{\eta} \eta_i^q) \cdot (\mathop{\mathbb{E}}_{\alpha} (\|z_i\|_2^2)^q) \right)^{1/q} \quad (\alpha, \eta \text{ independent}) \\
&= \left(m \cdot \max_i \mathop{\mathbb{E}}_{\alpha} (\|z_i\|_2^2)^q) \right)^{1/q} \\
&\le 2 \cdot \max_i \|\|z_i\|_2^2\|_q \ (m^{1/q} \le 2 \text{ by choice of } q) \\
&= 2 \cdot \max_i \|\alpha^T \tilde{U}_i \tilde{U}_i^T \alpha\|_q \\
&= 2 \cdot \max_i (d + \|\alpha^T \tilde{U}_i \tilde{U}_i^T \alpha - \mathbb{E}\, \alpha^T \tilde{U}_i \tilde{U}_i^T \alpha\|_q) \text{ (triangle inequality)}
\end{aligned} \tag{26}
$$

where $\tilde{U}_i$ is the matrix with $(\tilde{U}_i)_{k,j} = H_{i,k} \cdot U_{k,j}$. Of particular importance for us is the identity $\tilde{U}_i^T \tilde{U}_i = I$. Then by Eq. (26) and Theorem 8,

$$
\begin{aligned}
\|\max_i \eta_i \|z_i\|_2^2\|_q &\lesssim d + \sqrt{q} \cdot \|\tilde{U}_i \tilde{U}_i^T\|_F + q \cdot \|\tilde{U}_i \tilde{U}_i^T\| \\
&= d + \sqrt{qd} + q \\
&\le \frac{3}{2} \cdot (d+q) \text{ (AM-GM inequality)}
\end{aligned}
$$

so that

$$
R \lesssim \sqrt{\frac{p}{m}} \cdot \sqrt{d+q},
$$

which when combined with Eq. (25) gives

$$
\|\frac{1}{m} \sum_i \eta_i z_i z_i^T - I\|_p \lesssim \sqrt{\frac{p}{m} \cdot (d+q)} + \frac{p}{m} \cdot (d+q).
$$

Thus the OSE moment property is satisfied by our choices of m, p in the theorem statement. ■