

# Improved Approximation Algorithms for Large Matrices via Random Projections\*

Tamás Sarlós

Eötvös University and Computer and Automation Research Institute Hungarian Academy of Sciences  
Lágymányosi u. 11, Budapest, Hungary H-1111  
stamas@ilab.sztaki.hu

## Abstract

Recently several results appeared that show significant reduction in time for matrix multiplication, singular value decomposition as well as linear ( $\ell_2$ ) regression, all based on data dependent random sampling. Our key idea is that low dimensional embeddings can be used to eliminate data dependence and provide more versatile, linear time pass efficient matrix computation. Our main contribution is summarized as follows.

- Independent of the recent results of Har-Peled and of Deshpande and Vempala, one of the first – and to the best of our knowledge the most efficient – relative-error  $(1 + \epsilon) \|A - A_k\|_F$  approximation algorithms for the singular value decomposition of an  $m \times n$  matrix  $A$  with  $M$  non-zero entries that requires 2 passes over the data and runs in time

$$O\left(\left(M\frac{k}{\epsilon} + (n+m)\frac{k^2}{\epsilon^2}\right)\log\frac{1}{\delta}\right).$$

- The first  $o(nd^2)$  time  $(1 + \epsilon)$  relative-error approximation algorithm for  $n \times d$  linear ( $\ell_2$ ) regression.
- A matrix multiplication algorithm that easily applies to implicitly given matrices.

## 1 Introduction

This paper develops and analyzes fast approximation algorithms for fundamental linear algebra problems such as singular value decomposition (SVD), linear  $\ell_2$  regression and the computation of matrix products. Our motivation comes from the widespread use of these tools in data mining [10]. Prominent applications of low-rank matrix approximation by SVD include recommendation systems [25], information retrieval via Latent Semantic Indexing [13, 48], Kleinberg’s celebrated HITS algorithm for web search [43, 2], clustering [22, 47], and learning mixtures of distributions [42, 4] just to name a few. Classification can be solved by regularized regression [29] and text database querying by matrix-vector products [15].

While polynomial, all the three matrix operations mentioned above are computationally intensive when performed exactly. For example dense SVD methods require  $O(m^2n)$  time and  $O(mn)$  space on an  $m \times n$ ,  $m \leq n$ , matrix [36], both of which are prohibitively large even for moderate size data sets arising in current applications. Even for sparse data it is often the case that the input far exceeds the main memory and hence

---

\*The research was partially supported by the Inter-University Center for Telecommunications and Informatics (ETIK) and from the NKFP 2005 projects ASTOR and MOLINGV.

we generally restrict ourselves to the pass efficient “streaming” model of computation [38]. Here access to the input is limited to a constant number of sequential scans and RAM usage depends sublinearly on input size. Also note that sparse iterative SVD methods [36] alone are not suitable for streaming computation as their convergence speed is unknown a priori and thus generally they require too many passes over the input. Similarly, approximate SVD schemes based on the Lánczos or power method require  $\Omega(\log m)$  passes [44, 37].

Recently a large number of results appeared that prove bounds for non-uniform sampling to speed up approximate matrix operations [35, 48, 3, 23, 28, 24, 26, 49, 20]. These results provide error guarantees that depend on the Frobenius norm of the input matrices and hence may incur a large additive term. An exception among sampling based techniques is the sequel of results of Drineas et al. [29, 30, 31], Har-Peled [37], and Deshpande et al. [20, 21]. In the case of regression and singular value decomposition by using very special distributions for sampling they show that there exists a small subset of the input which contains a relative-error approximation. However, [29, 30, 31] give no advice for implementing the sampling procedure any faster than solving the original problem.

Low distortion embeddings also called “sketches” are known to outperform sampling in certain applications [14, 50]. Our key techniques to improve previous algorithms for singular value decomposition,  $\ell_2$  regression and matrix multiplication are Johnson-Lindenstrauss type embeddings [41]. Ironically, one of the first approximate singular value decomposition algorithms [48] was also embedding-based.

Our central result is a relative-error SVD algorithm (Theorem 14). Extending the work of [30, 35, 20] we show that if we form  $O(k/\epsilon)$  random linear combinations from the rows of  $A \in \mathbb{R}^{m \times n}$ , then the best rank- $k$  approximation within the row space generated by the random projection achieves relative-error  $(1 + \epsilon) \|A - A_k\|_F$  with constant probability. By repeating the procedure and choosing the best approximation we obtain the same error bound with high probability. The algorithm requires two passes over the data and runs in time  $O((Mk/\epsilon + (n + m)k^2/\epsilon^2) \log(1/\delta))$ . Independently of our work Har-Peled [37], and Deshpande and Vempala [21] also proved similar results. However, our procedure is faster in terms  $k$  than the more efficient of those, [21] that necessitates  $\Theta(k \log k)$  passes.

We also present the first  $o(nd^2)$  time  $(1 + \epsilon)$ -approximation algorithm for  $\ell_2$  regression with coefficient matrix  $A \in \mathbb{R}^{n \times d}$ ,  $n \geq d = \omega(\log n)$ , by replacing sampling in [29] with embeddings (Theorem 12). We offer novel analysis with improved bounds compared to [29], lowering the required number of reduced dimensions for sketches for example to  $O(d \log d/\epsilon)$  that matches to the enhanced bound of [30] for sampling. Plugging in the fast Johnson-Lindenstrauss transform of Ailon and Chazelle [5] allows us to obtain an  $O(nd \log n)$  time algorithm for  $\epsilon$  down to  $\omega((d \log d(d + \log^2 n))/(n \log n))$ .

As the simplest applications of our technique we derive algorithms for approximating matrix products whose time and space usage and error bound match to that of the column-row sampling based method [23] (Theorem 9). Unlike [23] our algorithms extend unchanged to approximating chain products and most importantly come with much stronger element-wise error bounds and work for approximating products of *unknown* matrices. The  $\ell_2$  regression and SVD results are based on precisely these properties. En route we also use embeddings to estimate the Frobenius norm of implicitly formed matrices akin to Freivalds’ technique [34] (Lemma 8).<sup>1</sup> This estimate then can be used as a black box tool to boost the probability of correctness.

The rest of the paper is organized as follows. After describing related results and basic facts about embeddings we give approximate matrix product and approximate error testing algorithms in Section 2. Based on these in Section 3 we give our new linear ( $\ell_2$ ) regression results. These results are used finally in Section 4 in our SVD algorithm.

---

<sup>1</sup>We remark that the earlier work of Ar et al. [52, 39] contains essentially the same result.

## 1.1 Comparison with previous results

Except for [48, 46], to the best of our knowledge, all prior work on speeding up matrix operations is based on sampling. Cohen and Lewis set up random walks to approximate non-negative matrix products [15]. In their ground-breaking paper Frieze, Kannan, and Vempala [35] showed that given matrix  $A$ , through non-uniform sampling it is possible to select a  $O(\text{poly}(k, \epsilon^{-1}))$  sized submatrix  $C$  of  $A$  such that i) with the help of  $C$  the description of a rank  $k$  matrix  $\widehat{A}_k$  can be computed in constant time and ii)  $\|A - \widehat{A}_k\|_F \leq \|A - A_k\|_F + \epsilon \|A\|_F$  holds with high probability, where  $A_k$  denotes best rank- $k$  approximation. Subsequent research based on sampling entire rows or columns with probability proportional to their squared Euclidean length resulted in more practical algorithms and refined analysis both for SVD [28, 24, 49, 20] and approximate matrix products [23]. Other line of research is based on random sparsification and quantization [3, 23, 8].

Although at first it may seem contradictory, approximate matrix product algorithms do not compute the final result itself, but reduce the problem two the product of two smaller (or sparser) matrices. If needed the latter can be more easily multiplied with the preferred exact method [36, 17, 16].

Returning to SVD, the best preliminary result with respect to the Frobenius norm was derived by Deshpande and Vempala [21] independently of our work, and shows that if we sample  $O(k^2 \log k + k/\epsilon)$  rows from  $A$  in  $O(k \log k)$  passes in an adaptive manner [20], then the best rank- $k$  approximation within the (row)space generated by the sample achieves relative-error  $(1 + \epsilon) \|A - A_k\|_F$  with probability at least  $3/4$ . That algorithm runs in time  $O(M(k^2 \log k + \frac{k}{\epsilon}) + (m+n)(k^2 \log k + \frac{k}{\epsilon})^2)$ , where  $M$  denotes the number of non-zeroes of  $A$ . While improving the running time, we also reduce the number of passes to 2. Historically the first relative-error SVD was given by Har-Peled [37], also independent of this work. Besides running in  $O(\log n)$  passes it is slower then the other two approaches as its running rime depends on the size of the input matrix  $mn$  instead of the number of non-zero entries  $M$ .

As the first of the two preliminary results for the least squares regression problem Drineas et al. [29] proved that if we sample  $r' = \text{poly}(\epsilon^{-1}, d)$  rows from  $A$  and  $b$  with the sampling probabilities satisfying certain criteria, then with high probability the optimum solution of the  $r'$ -by- $d$  downsampled problem gives an  $\epsilon$ -approximation to the original least squares problem. The same authors go a step further in [30] by showing that it is possible to construct a rank  $O(k \log k / \epsilon^2)$  matrix which approximates  $A$  to error  $(1 + \epsilon) \|A - A_k\|_F$  and its columns are expressible as linear combinations of a  $O(k \log k / \epsilon^2)$  sized subset of columns of  $A$ .

The crux of all column or row sampling proofs are the results that sampling provides good enough approximation for matrix products if the sampling probabilities are proportional to the column and row lengths of the matrices in question [23]. In fact uniform sampling is insufficient as [11] shows. In [29, 30, 31] these results are then applied to products arising from the singular value decomposition of the input. However, as noted, it is unknown whether the required nonuniform sampling probabilities can be computed any faster than the time required to solve the problem exactly.

In contrast, we observe that data independent random projections approximate dot products well, and hence are also capable of approximating matrix products within the same bounds as data dependent sampling. Our improved analysis for  $\ell_2$  regression directly exploits the low distortion of dot products.

## 1.2 Preliminaries

**Linear algebra and notation.** Let column vectors  $A_{(i)}$  and  $A^{(i)}$  denote  $i$ th row and column of matrix  $A \in \mathbb{R}^{m \times n}$ . Let  $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$  and  $\|A\|_F = \sqrt{\sum a_{ij}^2}$  denote the spectral and Frobenius norm of  $A$  respectively. The Singular Value Decomposition (SVD) of a rank  $\rho$  matrix  $A$  is given by  $A = U\Sigma V^T$

with  $U \in \mathbb{R}^{m \times \rho}$ ,  $\Sigma \in \mathbb{R}^{\rho \times \rho}$  and  $V \in \mathbb{R}^{n \times \rho}$ . By the Eckart-Young theorem the best rank- $k$  approximation of  $A$  with respect to both the Frobenius and spectral norms is  $A_k = U_k \Sigma_k V_k^T$ , where  $U_k \in \mathbb{R}^{m \times k}$  and  $V_k \in \mathbb{R}^{n \times k}$  contain the first  $k$  columns of  $U$  and  $V$  and the diagonal  $\Sigma_k \in \mathbb{R}^{k \times k}$  contains first  $k$  entries of  $\Sigma$ . For a subspace  $V \leq \mathbb{R}^m$  let  $\pi_V(A)$  denote the matrix formed by projecting every column of  $A$  to  $V$ . Similarly, let  $\Pi_W(A)$  stand for projecting each row of  $A$  to  $W \leq \mathbb{R}^n$  and let  $\Pi_{W,k}(A)$  denote the best rank- $k$  approximation of  $A$  with its rows in  $W$ , i.e.  $\Pi_{W,k}(A) = (\Pi_W(A))_k$ . Additionally, given matrix  $B$  let  $\text{colspan}(B) \leq \mathbb{R}^m$  and  $\text{rowspan}(B) \leq \mathbb{R}^n$  denote the subspaces generated by its column and rows, respectively and we use the simplified notation  $\pi_B(A)$  for  $\pi_{\text{colspan}(B)}(A)$  and  $\Pi_{B,k}(A)$  for  $\Pi_{\text{rowspan}(B),k}(A)$ . Furthermore let  $\sigma_i(A) = \Sigma_{ii}$  denote the  $i$ th singular value of  $A$  and let  $\sigma_{\min}(A) = \Sigma_{11}$  and  $\sigma_{\max}(A) = \Sigma_{\rho\rho}$ . The condition number of  $A$  is  $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$ . The Moore-Penrose generalized inverse of  $A$  can be expressed in terms of the SVD as  $A^+ = V\Sigma^{-1}U^T$ . For further linear algebra we refer the reader to [36].

**Random projections.** Johnson-Lindenstrauss's seminal paper [41] was followed by several variants and proofs of low-distortion embeddings [33, 40, 19]. Throughout this paper we will make extensive use of three flavors of  $\ell_2 \rightarrow \ell_2$  embeddings (Theorems 2 & 3, and Lemma 5); we list their properties now.

**Definition 1** A random matrix  $R \in \mathbb{R}^{k \times n}$  forms a Johnson-Lindenstrauss transform with parameters  $\epsilon, \delta, d$ , or  $JLT(\epsilon, \delta, d)$  for short, if there exists a function  $f$  that for any  $0 < \epsilon, \delta < 1$ , positive integer  $d$  and  $d$ -element subset  $V \subset \mathbb{R}^n$ , where  $k = \Omega(\frac{\log d}{\epsilon^2} f(\delta))$  with probability at least  $1 - \delta$  for all  $v \in V$  it holds that  $(1 - \epsilon) \|v\|_2^2 \leq \|Rv\|_2^2 \leq (1 + \epsilon) \|v\|_2^2$ .

**Theorem 2 (The Johnson-Lindenstrauss Lemma [19, 9])** Let  $0 < \epsilon, \delta < 1$  and  $\mathcal{S} = \frac{1}{\sqrt{k}}R \in \mathbb{R}^{k \times n}$  matrix such that the  $R_{ij} \sim N(0, 1)$  entries are independent standard normal random variables. If  $k = \Omega(\epsilon^{-2} \log d \log(1/\delta))$  then  $\mathcal{S}$  is a  $JLT(\epsilon, \delta, d)$ .

For practical applications the  $N(0, 1)$  entries can be replaced by random  $\pm 1$  variables [1, 9]. Recently Ailon and Chazelle showed [5] that a significantly sparser embedding matrix  $R$  suffices if inputs are preconditioned with a randomized Fast Fourier Transform and obtained a  $JLT(\epsilon, 2/3, d)$  which is faster to compute.

**Theorem 3 (The Fast  $\ell_2 \rightarrow \ell_2$  Johnson-Lindenstrauss Transform [5])** Let  $\mathcal{S} = \frac{1}{\sqrt{kn}}PH_nD$ , where  $D$  is an  $n \times n$  diagonal matrix with entries being independent uniformly random  $\pm 1$ ,  $H_n$  denotes the Hadamard-matrix of size  $n$  (w.l.o.g. we assume that  $n$  is a power of 2), and the entries of the  $k = O(\epsilon^{-2} \log d) \times n$  matrix  $P$  are i.i.d.  $N(0, q^{-1})$  with probability  $q$ , and 0 otherwise, where  $N = \max\{n, d\}$  and  $q = \min\{\Theta(n^{-1} \log^2 N), 1\}$ . Let  $\epsilon_0$  be an absolute constant. Then for any  $\epsilon \leq \epsilon_0$  and  $V \subset \mathbb{R}^n$ ,  $|V| = d$ , with probability at least  $2/3$  the following two events occur:

- For all  $v \in V$  it holds that  $(1 - \epsilon) \|v\|_2^2 \leq \|\mathcal{S}v\|_2^2 \leq (1 + \epsilon) \|v\|_2^2$ .
- For all  $x \in \mathbb{R}^n$  computing  $\mathcal{S}x$  takes  $O(n \log n + \epsilon^{-2} \log^2 N \log d)$  time.

Now let us consider the dot product  $\langle \mathcal{S}u, \mathcal{S}v \rangle$  for  $u, v \in V$ . By the parallelogram rule it is easy to see [9, 48] that if  $\mathcal{S}$  distorts squared norms by factor of at most  $1 \pm \epsilon$  and the set  $V$  contains unit length vectors only then  $|\langle \mathcal{S}u, \mathcal{S}v \rangle - \langle u, v \rangle| \leq \epsilon$ . If  $u = 0$  then trivially  $\langle \mathcal{S}0, \mathcal{S}v \rangle = \langle 0, v \rangle = 0$  for all  $v$ . If  $u \neq 0$  and  $v \neq 0$  then by linearity  $\langle \mathcal{S}u, \mathcal{S}v \rangle = \|u\|_2 \|v\|_2 \left\langle \mathcal{S} \frac{u}{\|u\|_2}, \mathcal{S} \frac{v}{\|v\|_2} \right\rangle$  and therefore we also have the following stronger corollary, to which we will often refer to.

**Corollary 4** If  $\mathcal{S}$  is a  $JLT(\epsilon, \delta, d)$ ,  $0 < \epsilon \leq 1$ , then for any  $V \subset \mathbb{R}^n$ ,  $|V| = d$  with probability at least  $1 - \delta$  for all  $u, v \in V$  it holds that  $\langle u, v \rangle - \epsilon \|u\|_2 \|v\|_2 \leq \langle \mathcal{S}u, \mathcal{S}v \rangle \leq \langle u, v \rangle + \epsilon \|u\|_2 \|v\|_2$ .

The last ingredient of our proofs was presented by Alon, Matias, and Szegedy in their seminal paper [7].

**Lemma 5 (Tug-of-war sketch, [7, 6])** *Let  $0 < \epsilon \leq 1$  and  $S = \epsilon R \in \mathbb{R}^{\epsilon^{-2} \times n}$  be a random matrix such that rows of  $R$  are independent and each row consists of a vector of four-wise independent zero-mean  $\{-1, +1\}$  random variables. Then for any  $x, y \in \mathbb{R}^n$  we have that  $\mathbf{E}(\langle Sx, Sy \rangle) = \langle x, y \rangle$  and  $\mathbf{Var}(\langle Sx, Sy \rangle) \leq 2\epsilon^2 \|x\|_2^2 \|y\|_2^2$ .*

## 2 Matrix multiplication

In this section we demonstrate the versatility of sketches by devising pass efficient algorithms for approximating matrix products. The algorithms to be presented are based on the following simple, but powerful observation.

**Lemma 6** *Let  $A \in \mathbb{R}^{m \times n}$ ,  $B \in \mathbb{R}^{n \times p}$ .*

- *If  $S$  is a  $JLT(\epsilon, \delta, (m+p))$ , then  $\Pr(\|AB - AS^T SB\|_F \leq \epsilon \|A\|_F \|B\|_F) \geq 1 - \delta$ .*
- *If  $S$  is  $\epsilon^{-2} \times n$  tug-of-war random matrix then  $\mathbf{E}(AS^T SB) = AB$  and  $\mathbf{E}\left(\|AB - AS^T SB\|_F^2\right) \leq 2\epsilon^2 \|A\|_F^2 \|B\|_F^2$ .*

PROOF: Set  $a_k = A_{(k)}$  and  $b_k = B_{(k)}$ . Note that  $(AS^T)_{(i)} = Sa_i$  and  $(SB)^{(j)} = Sb_j$  and thus  $Y_{ij} = (AB)_{ij} - (AS^T SB)_{ij} = \langle a_i, b_j \rangle - \langle Sa_i, Sb_j \rangle$ .

For the first claim let  $V = \{a_1, \dots, a_m, b_1, \dots, b_p\}$ . Then by Corollary 4 with probability at least  $1 - \delta$  for all  $i, j$  we have that  $|Y_{ij}| \leq \epsilon \|a_i\|_2 \|b_j\|_2$ , and thus

$$\|AB - AS^T SB\|_F^2 = \sum_{i,j} Y_{ij}^2 \leq \sum_{i,j} \epsilon^2 \|a_i\|_2^2 \|b_j\|_2^2 = \epsilon^2 \|A\|_F^2 \|B\|_F^2.$$

For the second statement observe that by Lemma 5 the expected value  $\mathbf{E}(Y_{ij}) = 0$  and consequently  $\mathbf{E}(Y_{ij}^2) = \mathbf{Var}(\langle Sa_i, Sb_j \rangle) \leq 2\epsilon^2 \|a_i\|_2^2 \|b_j\|_2^2$ .  $\square$

Combining the first statement of Lemma 6 with Lemma 2 immediately gives us a one pass algorithm which uses  $O(\epsilon^{-2} \log(m+p) \log(1/\delta)(m+p))$  space and  $O(\epsilon^{-2} \log(m+p) \log(1/\delta)M)$  time to output matrices  $\hat{A} = AS^T \in \mathbb{R}^{m \times O(\epsilon^{-2} \log(m+p) \log(1/\delta))}$  and  $\hat{B} = SB \in \mathbb{R}^{O(\epsilon^{-2} \log(m+p) \log(1/\delta)) \times p}$  such that  $\Pr(\|AB - \hat{A}\hat{B}\|_F \leq \epsilon \|A\|_F \|B\|_F) \geq 1 - \delta$ , where  $M$  denotes the number of non-zero entries in  $A$  and  $B$  altogether. The complexity of the procedure is a factor  $\log(m+p)$  higher than that of the column-row sampling approach [23].

Next, we remove the  $\log(m+p)$  factor by proving the same high probability bound using the second claim of Lemma 6. In fact, our method is more general and it can be used to turn a large class of matrix approximation algorithms having low error in the Frobenius norm with constant probability to an algorithm having the same low error with high probability.

In the case of matrix product by Lemma 6 and Markov's inequality we have

**Fact 7** *Given  $0 < \delta < 1$  let us instantiate  $t = \log 1/\delta$  independent copies of the tug-of-war matrix  $S_i$ . Then  $\Pr(\min_{i=1 \dots t} \|AB - AS_i^T S_i B\|_F \leq 2\epsilon \|A\|_F \|B\|_F) \geq 1 - \delta$ .*

However, it is non-trivial to choose the best  $\mathcal{S}_i$ . Computing  $\|AB - AS_i^T \mathcal{S}_i B\|_F^2$  exactly requires i) one full  $AB$  matrix multiplication (which we are trying to approximate) ii) at least  $\Omega(mp)$  space/time, which is way too high for us. To overcome this, we will apply the tug-of-war trick once more to approximate squared Frobenius norms and hence pick (almost) the best  $AS_i^T \mathcal{S}_i B$ . Similarly to Freivalds' technique for *checking* matrix products our norm estimation method requires a few extra matrix-vector products only [34] and was motivated by Lemma 4 of [18].

**Lemma 8** *Let  $C$  be an  $m \times n$  matrix,  $0 < \lambda < 1$ , and  $\mathcal{Q}$  a  $\lambda^{-2} \times n$  tug-of-war random matrix as in Lemma 5. Define  $X = \|C\mathcal{Q}^T\|_F^2$ . Then  $\mathbf{E}(X) = \|C\|_F^2$  and  $\mathbf{Var}(X) \leq 2\lambda^2 \|C\|_F^4$ .*

PROOF: We proceed similarly to Lemma 6. Let  $\mathcal{Q} = \lambda R$  and  $r_i$  denote the  $i$ th row of the unscaled tug-of-war matrix  $R$ . Set  $Y_i = \|Cr_i\|_2^2$  and  $c_j = C_{(j)}$ . Observe that  $\|C\mathcal{Q}^T\|_F^2 = \sum_{i=1}^{1/\lambda^2} \lambda^2 \|Cr_i\|_2^2$  and hence it is enough to show that  $\mathbf{E}(Y_i) = \|C\|_F^2$  and  $\mathbf{Var}(Y_i) \leq 2\|C\|_F^4$  hold. Using Lemma 5

$$\mathbf{E}\left(\|Cr_i\|_2^2\right) = \mathbf{E}\left(\sum_j \langle c_j, r_i \rangle^2\right) = \sum_j \mathbf{E}\left(\langle c_j, r_i \rangle^2\right) = \sum_j \langle c_j, c_j \rangle = \|C\|_F^2.$$

By the Cauchy-Schwartz inequality and  $\mathbf{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}^2(X)$  it follows from Lemma 5 that for arbitrary vectors  $x, y$  we have that  $\mathbf{E}(\langle x, r_i \rangle \langle y, r_i \rangle) \leq 3\|x\|_2 \|y\|_2$ . Hence

$$\begin{aligned} \mathbf{E}(Y_i^2) &= \mathbf{E}\left(\left(\sum_j \langle c_j, r_i \rangle^2\right)^2\right) = \sum_{j,k} \mathbf{E}\left(\langle c_j, r_i \rangle \langle c_k, r_i \rangle\right)^2 \\ &\leq 3 \sum_{j,k} \|c_j\|_2^2 \|c_k\|_2^2 = 3 \left(\|C\|_F^2\right)^2 = 3\|C\|_F^4. \end{aligned}$$

Thus  $\mathbf{Var}(Y_i) = \mathbf{E}(Y_i^2) - \mathbf{E}^2(Y_i) \leq 2\|C\|_F^4$ .  $\square$

Observing that the test outlined above trivially extends to multivariate matrix polynomials, we are ready to present our algorithm.

---

**Algorithm 1** Approximate product of  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{n \times p}$  matrices by tug-of-war sketches

---

- 1: **for**  $i = 1, \dots, \log 1/\delta$  **do**
  - 2:   Pick  $\mathcal{S}_i \in \mathbb{R}^{1/\epsilon^2 \times n}$  random tug-of-war matrices as in Lemma 5.
  - 3: **for**  $i = 1, \dots, \log 1/\delta$ ,  $j = 1, \dots, 2(\log 1/\delta + \log \log 1/\delta)$  **do**
  - 4:   Pick  $\mathcal{Q}_{i,j} \in \mathbb{R}^{16 \times p}$  random tug-of-war matrices as in Lemma 5.
  - 5:   Compute  $\mathcal{S}_i B$ ,  $AS_i^T$ .
  - 6:   Compute  $B\mathcal{Q}_{i,j}^T$  and then  $X_{i,j} = A(B\mathcal{Q}_{i,j}^T)$ .
  - 7:   Compute  $(\mathcal{S}_i B)\mathcal{Q}_{i,j}^T$  and then  $\hat{X}_{i,j} = (AS_i^T)(\mathcal{S}_i B\mathcal{Q}_{i,j}^T)$ .
  - 8:   Let  $y_{i,j} = \|X_{i,j} - \hat{X}_{i,j}\|_F^2$ .
  - 9:   Let  $z_i = \text{median}_j y_{i,j}$ .
  - 10:   Choose  $i^*$  with minimal  $z_i$ .
- 

**Theorem 9** *For Algorithm 1 we have that  $\Pr(\|AB - AS_{i^*}^T \mathcal{S}_{i^*} B\|_F \leq \sqrt{12}\epsilon \|A\|_F \|B\|_F) \geq 1 - 2\delta$  and  $\mathbf{E}(AB - AS_{i^*}^T \mathcal{S}_{i^*} B) = \mathbf{0}$ . If  $M$  denotes the total number of non-zeroes in  $A$  and  $B$  then the algorithm runs in at most two passes in  $O((m+M)(\epsilon^{-2} \log 1/\delta + (\log 1/\delta)^2))$  time and uses  $O((m+n+p)(\epsilon^{-2} \log 1/\delta + (\log 1/\delta)^2))$  space and requires at most two passes over the data.*

With Lemmas 6 and 8 at hand the proof is a routine application of Chebyshev’s inequality and hence it is deferred to Appendix B. Comparing Algorithm 1 to column-row sampling [23] we observe that their proven bounds are equivalent, but the embedding based Algorithm 1 extends to multiple term matrix products unmodified unlike column-row sampling. For further discussion we refer the reader to Appendix B.

### 3 The $\ell_2$ regression

In this section we present an approximation algorithm for the least squares regression problem, i.e. given an  $n$ -by- $d$ ,  $n > d$ , matrix  $A$  of reals and a  $d$  dimensional real vector  $b$  we wish to obtain  $x_{opt} = A^+b$  minimizing  $\|Ax - b\|_2$ . Recall that the preliminary results proven by Drineas et al. [29, 30] show that if we sample  $r' = \text{poly}(\epsilon^{-1}, d)$  rows from  $A$  and  $b$  with the sampling probabilities satisfying certain criteria, then with high probability the optimum solution of the  $r'$ -by- $d$  downsampled problem gives an  $\epsilon$ -approximation to the original least squares problem. However, it is unknown whether the required nonuniform sampling probabilities can be computed any faster than the  $O(nd^2)$  time required to solve the problem exactly.

Firstly, we observe that all the claims and proofs of [29] carry through unmodified if we project the input by forming  $r = O(r')$  random linear combinations of  $A$  and  $b$ ’s rows as sketches provide good enough approximation for matrix products (details are omitted). Secondly, we independently analyze the random projection based method and significantly lower the bounds for the required reduced dimension  $r$  for all the main statements of [29], i.e. we improve it from  $r' = O(d^2/\epsilon^4)$  to  $r = O(\epsilon^{-2})$ , from  $r' = O(d^2/\epsilon^2)$  to  $r = O(\epsilon^{-1}d \log d)$ , and from  $r' = O(d^2/\epsilon^2)$  to  $r = O(\epsilon^{-2}d \log d)$ . These bounds for sketching are on par even with those obtainable by a more careful reading of the recent enhanced sampling proofs in [30]. Thirdly, plugging in the Fast Johnson-Lindenstrauss Transform (FJLT, Theorem 3) for the random projection allows us to obtain an  $O(nd \log n)$  time algorithm. We remark that for  $d = O(\log n)$  the exact solution is efficient itself. In what follows we state the input parameters  $\tilde{\epsilon}$  and  $\tilde{d}$  of the (F)JLT implicitly as  $r = \Omega(\tilde{\epsilon}^{-2} \cdot \log \tilde{d})$  for easier comparison with [29]

The Johnson-Lindenstrauss Lemma states that  $k$  vectors from  $\mathbb{R}^m$  can be embedded into  $O(\log(k)/\epsilon^2)$  dimensions such that the length of each vector is preserved up to a  $1 + \epsilon$  factor (see Section 1.2). It is easy to see that given a  $k$  dimensional subspace  $V$ , embedding it into  $O(k^2 \log(k)/\epsilon^2)$  dimensions preserves the length of all vectors from  $V$ . However, it follows from a lemma of Feige and Ofek [32] based on putting a grid on the unit sphere that mere  $O(k/\epsilon^2)$  dimensions are sufficient. We remark that the same lemma and grid construction also appeared in [12, 51]; [45] contains a weaker form. Even though the dimension of the target subspace is significantly higher than  $k$ , the embedding will still turn out to be useful as it can be constructed without knowing the subspace  $V$ .

**Lemma 10 ([32], see also [8] for a restated proof)** *Let  $0 < \epsilon_0, T = \{x : x \in \frac{\epsilon_0}{\sqrt{k}}\mathbb{Z}^k, \|x\|_2 \leq 1\}$  and  $C \in \mathbb{R}^{k \times k}$ . The number of vectors in  $T$  is at most  $e^{k \ln(9/\epsilon_0)}$ . If for all  $x, y \in T$  we have  $|x^T C y| \leq \epsilon$  then for all unit vector  $x \in \mathbb{R}^k$ , we have  $|x^T C x| \leq \frac{\epsilon}{(1-\epsilon_0)^2}$ .*

**Corollary 11** *Let  $0 < \epsilon, \delta < 1$  and  $\mathcal{S}$  be a Johnson-Lindenstrauss transform to  $O(k/\epsilon^2 \cdot f(\delta))$  dimensions for some function  $f$ .*

- (Subspace JL Lemma) *If  $\mathcal{S}$  is a JLT from  $\mathbb{R}^m$  and  $V$  is an arbitrary  $k$  dimensional subspace of  $\mathbb{R}^m$  then*

$$\Pr \left( \forall v \in V : (1 - \epsilon) \|v\|_2^2 \leq \|\mathcal{S}v\|_2 \leq (1 + \epsilon) \|v\|_2^2 \right) \geq 1 - \delta,$$

or stated otherwise, if  $U \in \mathbb{R}^{m \times k}$ ,  $m \geq k$ , is a unitary matrix then

$$\Pr(\forall i \in [1..k] : |1 - \sigma_i(SU)| \leq \epsilon) \geq 1 - \delta.$$

- (Weak) spectral bound for approximate matrix products. If  $A \in \mathbb{R}^{m \times k}$ ,  $B \in \mathbb{R}^{k \times p}$  and  $S$  is a JLT from  $\mathbb{R}^k$  then

$$\Pr(\|AS^T SB - AB\|_2 \leq \epsilon \|A\|_2 \|B\|_2) \geq 1 - \delta.$$

PROOF: For the first statement let  $U \in \mathbb{R}^{m \times k}$  be an orthonormal basis of the subspace  $V$ . Set  $C = U^T S^T S U - I_k$ ,  $T' = \{Ux : x \in T\}$ , and  $\epsilon_0 = 1/2$ . Applying Corollary 4 to the set  $T'$ , from  $|T'| = O(\exp(k))$  it follows that with high probability we have  $|\langle SUx, SUy \rangle - \langle x, y \rangle| = |x^T C y| \leq \epsilon/4$  for every  $x, y \in T$ . Hence by Lemma 10 w.h.p. for all unit vector  $x \in \mathbb{R}^k$  it holds that  $|x^T C x| \leq \epsilon$  and thus  $|\|SUx\|_2^2 - \|Ux\|_2^2| \leq \epsilon$  proving the first claim. The second statement is just a reformulation of the fact that w.h.p. for any unit length vector  $x \in \mathbb{R}^k$  it holds that  $1 - \epsilon \leq \|SUx\|_2 \leq 1 + \epsilon/2$ .

For the last statement set  $U = I_k$  and thus  $C = S^T S - I_k$  and observe that  $\|AS^T SB - AB\|_2 = \|A(S^T S - I_k)B\|_2 \leq \|A\|_2 \|B\|_2 \|C\|_2$ . As  $C$  is symmetric we have that  $\|C\|_2 = \max_{\|x\|_2=1} x^T C x \leq \epsilon$  concluding the proof.  $\square$

**Theorem 12** Suppose  $A \in \mathbb{R}^{n \times d}$ ,  $b \in \mathbb{R}^n$ . Let  $\mathcal{Z} = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 = \|b - Ax_{opt}\|_2$ , where  $x_{opt} = A^+ b$  is a minimizer of the above formula. Let  $0 < \epsilon < 1$  and  $\mathcal{S}$  be a Johnson-Lindenstrauss Transform from  $\mathbb{R}^n$  to  $\mathbb{R}^r$  and  $\tilde{\mathcal{Z}} = \min_{x \in \mathbb{R}^d} \|\mathcal{S}b - \mathcal{S}Ax\|_2 = \|\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt}\|_2$ , where  $\tilde{x}_{opt} = (\mathcal{S}A)^+ \mathcal{S}b$ .

- If  $r = \Omega(\epsilon^{-2})$  then with probability at least  $2/3$

$$\tilde{\mathcal{Z}} \leq (1 + \epsilon)\mathcal{Z}. \quad (1)$$

- If  $r = \Omega(\epsilon^{-1} d \cdot \log d)$  then with prob. at least  $1/3$

$$\|b - A\tilde{x}_{opt}\|_2 \leq (1 + \epsilon)\mathcal{Z}. \quad (2)$$

- If  $r = \Omega(\epsilon^{-2} d \cdot \log d)$  then with prob. at least  $1/3$

$$\|x_{opt} - \tilde{x}_{opt}\|_2 \leq \frac{\epsilon}{\sigma_{\min}(A)} \mathcal{Z}. \quad (3)$$

Furthermore computing  $\tilde{x}_{opt}$  by the Fast Johnson-Lindenstrauss Transform takes  $O(nd \log n + d^2(d + \log^2 n) \log d \epsilon^{-p})$  time with  $p = 1$  in the case of (2) and  $p = 2$  in the case of (3), thus for  $d = \omega(\log n)$  we achieve (2) in  $o(nd^2)$  time if  $\epsilon = \omega(\log d(d + \log^2 n)/n)$ .

PROOF:[Inequality 1] Applying the JLT to the single vector  $b - Ax_{opt}$ , by linearity we immediately obtain

$$\tilde{\mathcal{Z}} = \|\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt}\|_2 \leq \|\mathcal{S}(b - Ax_{opt})\|_2 \leq (1 + \epsilon) \|b - Ax_{opt}\|_2 = (1 + \epsilon)\mathcal{Z}.$$

[Inequality 2] Let  $A = U\Sigma V^T$  be the SVD of  $A$  and  $\rho = \text{rank}(A) \leq d$ . Additionally set  $\alpha, \beta \in \mathbb{R}^\rho$  and  $w \in \mathbb{R}^n$  such that  $Ax_{opt} = U\alpha$ ,  $b = Ax_{opt} + w$  and  $A\tilde{x}_{opt} - Ax_{opt} = U\beta$  hold. Thus  $w$  is orthogonal to  $\text{colspan}(U)$  and  $\|w\|_2 = \mathcal{Z} = \|b - Ax_{opt}\|_2$  and we have that

$$\|b - A\tilde{x}_{opt}\|_2^2 = \|w - U\beta\|_2^2 = \mathcal{Z}^2 + \|\beta\|_2^2. \quad (4)$$

To upper bound  $\|\beta\|_2^2$ , recall that  $\pi(\cdot)$  denotes the column projection operator defined in Section 1.2 and observe that  $SU(\alpha+\beta) = SA\tilde{x}_{opt} = SA(SA)^+Sb = \pi_{SA}(Sb) = \pi_{SU}(Sb)$  as  $\text{colspan}(SU) = \text{colspan}(SA)$ . From  $\pi_{SU}(Sb) = \pi_{SU}(S(U\alpha + w)) = SU\alpha + \pi_{SU}(Sw)$  it follows that  $SU\beta = \pi_{SU}(Sw)$  and hence

$$U^T S^T \cdot SU\beta = U^T S^T \cdot Sw. \quad (5)$$

Now, as  $r = \Omega(\rho)$  setting  $\epsilon'$  to  $1 - 1/\sqrt[4]{2}$  in Corollary 11 with probability at least  $2/3$  gives us  $\sigma_i(U^T S^T SU) = \sigma_i^2(SU) \geq 1/\sqrt{2}$  and thus

$$\|\beta\|_2^2 / 2 \leq \|U^T S^T SU\beta\|_2^2 = \|U^T S^T Sw\|_2^2.$$

Applying the first statement of Lemma 6 with  $\epsilon' = \sqrt{\epsilon/d}$  to  $U^T$  and  $w$ , from  $U^T w = \mathbf{0}$  it follows that

$$\|U^T S^T Sw\|_2^2 \leq \epsilon \|w\|_2^2 = \epsilon \mathcal{Z}^2$$

holds with probability at least  $2/3$ . By the union bound with probability at least  $1/3$  we arrive at

$$\|\beta\|_2^2 \leq 2\epsilon \mathcal{Z}^2.$$

Combining the latter with equation (4) we conclude the proof of the second claim by observing that

$$\|b - A\tilde{x}_{opt}\|_2 \leq \sqrt{1 + 2\epsilon} \mathcal{Z} \leq (1 + \epsilon) \mathcal{Z}.$$

[Inequality 3] Reusing the previous proof with  $\epsilon \leftarrow \epsilon^2$  we have  $\|\beta\|_2^2 \leq 2\epsilon^2 \mathcal{Z}^2$  and  $A(x_{opt} - \tilde{x}_{opt}) = U\beta$ . Thus  $(\Sigma V^T)(x_{opt} - \tilde{x}_{opt}) = \beta$  since  $U$  is orthogonal. Note that for all  $1 \leq i \leq \rho$  we have  $\sigma_i(\Sigma V^T) = \sigma_i(A) > 0$ .

For bounding  $\|x_{opt} - \tilde{x}_{opt}\|_2$  it is crucial to recall that by  $x_{opt} = A^+b = V\Sigma^{-1}U^Tb$  is a linear combination of columns of  $V$  (the right singular vectors) and hence  $x_{opt}$  lies in the row space of  $A$  denoted by  $\text{rowspan}(A)$ . Similarly  $\tilde{x}_{opt}$  lies in  $\text{rowspan}(SA)$ , which in turn is contained in  $\text{rowspan}(A)$ , since the rows of  $SA$  are formed by random linear combinations of rows of  $A$ . Consequently  $x_{opt} - \tilde{x}_{opt} = \sum_{i=1}^{\rho} \eta_i v_i$  for some  $\eta \in \mathbb{R}^{\rho}$ ,  $\|\eta\|_2 = \|x_{opt} - \tilde{x}_{opt}\|_2$ , and hence  $\sum_{i=1}^{\rho} \sigma_i^2 \eta_i^2 = \|\beta\|_2^2$ . We establish the third claim by additionally observing that

$$\sigma_{\min}(A) \|x_{opt} - \tilde{x}_{opt}\|_2 = \sigma_{\min}(\Sigma V^T) \|\eta\|_2 = \sqrt{\sum_{i=1}^{\rho} \sigma_{\rho}^2 \eta_i^2} \leq \sqrt{\sum_{i=1}^{\rho} \sigma_i^2 \eta_i^2} = \|\beta\|_2 \leq 2\epsilon \mathcal{Z}. \quad \square$$

**Remark.** Although Theorem 12 guarantees only a constant probability of success, it is easy to see that by repeating the projection  $\log(1/\delta)$  times inequalities (1-3) hold with probability at least  $1 - \delta$  for the outcome  $\tilde{x}_{opt}^*$  with minimal  $\tilde{\mathcal{Z}}$  or  $\mathcal{Z}$  value, respectively.

If  $\sqrt{\|b\|_2^2 - \mathcal{Z}^2} \geq \gamma \|b\|_2$  for some  $0 < \gamma \leq 1$ , then with any  $r$  for (3), with probability at least  $1/3$  we have that  $\|x_{opt} - \tilde{x}_{opt}\|_2 \leq 5\epsilon \left( \kappa(A) \sqrt{\gamma^{-2} - 1} \right) \|x_{opt}\|_2$  since it follows from (3) as the proof of inequality (3.16) in [29] shows.

We conclude this section by observing that the proof of inequalities (2) and (3) works unchanged for any matrix  $S$  such that  $|1 - \sigma_i^2(SU)| = o(1)$  and  $US^T Sw \approx U^T w$ . Thus combining the above with Rudelson's and Vershynin's proof of Theorem 1.1 in [49] for bounding the singular values and Lemma 8 in appendix A.2 of [23] for bounding the norm of the approximate matrix product we have the following claim for sampling  $\ell_2$  regression.

**Claim 13** Let  $r > 0$  and for all  $1 \leq i \leq n$  set  $p_i = \frac{\|U_{(i)}\|_2^2}{\|U\|_F^2}$ . Let  $\mathcal{S} \in \mathbb{R}^{r \times n}$  be a row-sampling matrix such that  $\Pr\left(S_{(j)} = \frac{e_i}{\sqrt{r p_i}}\right) = p_i$  for all  $1 \leq j \leq r$ , where  $e_i$  denotes the  $i$ th unit vector. Then for any  $0 < \epsilon \leq 1$  inequalities (2) and (3) also hold with probability at least  $1/3$  if  $r = \Omega(d \log d + d\epsilon^{-1})$  and  $r = \Omega(d \log d + d\epsilon^{-2})$ , respectively.

We observe that the FJLT never requires more dimensions for (1) than the  $\Omega(d \log d + \epsilon^{-2})$  obtainable for sampling if one defers the square root to the very end of the proofs in [29, 30]. The latter modification also yields  $r = O(d \log(d)/\epsilon)$  for the sampling version of inequality (2), which matches to the bound of the FJLT. However Claim 13 asks for even less in the case of sampling.

## 4 Relative-error SVD

In this section we present a relative-error approximate Singular Value Decomposition algorithm, i.e. given an  $m \times n$ ,  $m < n$ , matrix of  $A$  of reals we wish to obtain  $A_k = U_k \Sigma_k V_k^T$ , minimizing  $\|A - X_k\|_F$  among the rank- $k$  matrices  $X_k$ .

Adapting the proofs of [30] we show that if we form  $O(k/\epsilon)$  random linear combinations of rows of  $A$  then the best rank- $k$  approximation within the (row)space generated by the random projection achieves relative-error  $(1+\epsilon) \|A - A_k\|_F$  with constant probability, which we then boost to arbitrary high probability. The resulting algorithm runs in time  $O((Mk/\epsilon + (n+m)k^2/\epsilon^2) \log(1/\delta))$ , where  $M$  denotes the number of non-zeroes in  $A$ .

**Theorem 14** Let  $A \in \mathbb{R}^{m \times n}$  and recall that  $\Pi_{(\cdot)}$  denote the row projection operators defined in Section 1.2. If  $0 < \epsilon \leq 1$  and  $\mathcal{S}$  is an  $r$ -by- $n$  Johnson-Lindenstrauss matrix with i.i.d. zero-mean  $\pm 1$  entries and  $r = \Theta(k/\epsilon)$  then with probability at least  $1/2$  it holds that

$$\|A - \Pi_{\mathcal{S}A,k}(A)\|_F \leq (1 + \epsilon) \|A - A_k\|_F.$$

Computing the singular vectors spanning  $\Pi_{\mathcal{S}A,k}(A)$  in two passes over the data requires  $O(Mr + (m+n)r^2)$  time and  $O((m+n)r^2)$  space, where  $M$  denotes the number of non-zeroes in  $A$ .

PROOF: We will proceed similarly to the proof of Theorem 1 in [30] but in a more concise way. Let  $A = U \Sigma V^T$  be the SVD of  $A$  and  $\rho = \text{rank}(A)$ . Slightly abusing the notation let matrices  $U_k$  and  $U_{\rho-k}$  contain the first  $k$  and last  $\rho - k$  columns of  $U$  respectively. Observe that for all vectors  $x \in \mathbb{R}^k$  and  $y \in \mathbb{R}^{\rho-k}$  by the Pythagorean theorem it holds that  $\|U_k x + U_{\rho-k} y\|_2^2 = \|U_k x\|_2^2 + \|U_{\rho-k} y\|_2^2$ . Hence we have that  $\|U_k X + U_{\rho-k} Y\|_F^2 = \|U_k X\|_F^2 + \|U_{\rho-k} Y\|_F^2$  for all matrices  $X \in \mathbb{R}^{k \times n}$  and  $Y \in \mathbb{R}^{(\rho-k) \times n}$ .

Now, let  $D \in \mathbb{R}^{m \times n}$  be any matrix of the form  $D = U_k \cdot C \cdot \mathcal{S}A$  with arbitrary  $C \in \mathbb{R}^{k \times r}$ . It follows that

$$\|A - \Pi_{\mathcal{S}A,k}(A)\|_F^2 \leq \|A - D\|_F^2 \tag{6}$$

as  $\text{rank}(D) \leq k$  and the rows of  $D$  lie in the row space of  $\mathcal{S}A$  and  $\Pi_{\mathcal{S}A,k}(A)$  is the best rank- $k$  approximation of  $A$  from the row space of  $\mathcal{S}A$ . Furthermore, note that

$$\|A - D\|_F^2 = \|A - A_k\|_F^2 + \|A_k - D\|_F^2 \tag{7}$$

since  $A - A_k$  lies in the column space of  $U_{\rho-k}$  and  $A_k - D$  lies in the column space of  $U_k$ .

We set  $D = A_k(\mathcal{S}A_k)^+(\mathcal{S}A)$ . To complete the proof it is sufficient to show that with probability at least  $1/2$  we have  $\|A_k - D\|_F^2 \leq 2\epsilon \|A - A_k\|_F^2$ , since combining the latter with equations (6-7) immediately gives us

$$\|A - \Pi_{\mathcal{S}A,k}(A)\|_F \leq \sqrt{(1 + 2\epsilon) \|A - A_k\|_F^2} \leq (1 + \epsilon) \|A - A_k\|_F.$$

Recall that  $Y^{(j)}$  denotes the  $j$ th column of matrix  $Y$  and let us consider the regressions  $A^{(j)} \approx A_k x_j$  for  $j = 1, \dots, n$ . Note that the best approximation of  $A^{(j)}$  from  $A_k$  is  $\pi_{A_k}(A^{(j)}) = A_k^{(j)}$  and hence it follows as equations (4-5) in the proof of Theorem 12 that there exists vectors  $\beta_1, \dots, \beta_n \in \mathbb{R}^k$  and  $w_1, \dots, w_n \in \mathbb{R}^m$  orthogonal to  $\text{colspan}(U_k)$  such that

$$\begin{aligned} \forall j \in \{1, \dots, n\} : \|w_j\|_2^2 &= \|A^{(j)} - A_k^{(j)}\|_2^2 \\ \forall j \in \{1, \dots, n\} : U_k^T \mathcal{S}^T \mathcal{S} U_k \beta_j &= U_k^T \mathcal{S}^T \mathcal{S} w_j, \text{ and} \\ \sum_{j=1}^n \|\beta_j\|_2^2 &= \|A_k - A_k(\mathcal{S}A_k)^+(\mathcal{S}A)\|_F^2 \end{aligned} \quad (8)$$

From  $r = \Omega(k)$  and Corollary 11 we have  $\|\beta_j\|_2^2/2 \leq \|U_k^T \mathcal{S}^T \mathcal{S} U_k \beta_j\|_2^2$  with probability  $3/4$  for all  $j$  as before. Observing that  $\mathcal{S}$  is a tug-of-war matrix as well and applying the second statement of Lemma 6 with  $\epsilon' = \sqrt{\epsilon/k}$  to  $U_k^T$  and  $w_j$  from  $U_k^T w_j = \mathbf{0}$  it follows that

$$\mathbf{E} \left( \sum_{j=1}^n \|U_k^T \mathcal{S}^T \mathcal{S} w_j\|_2^2 \right) = \sum_{j=1}^n \mathbf{E} \left( \|U_k^T \mathcal{S}^T \mathcal{S} w_j\|_2^2 \right) \leq \sum_{j=1}^n \epsilon \|w_j\|_2^2 = \epsilon \sum_{j=1}^n \|A^{(j)} - A_k^{(j)}\|_2^2$$

Thus by Markov's inequality and the union bound we have that  $\sum_{j=1}^n \|\beta_j\|_2^2 \leq 8 \|A - A_k\|_F^2$  holds with probability at least  $1/2$ . Combining the latter with equation (8) and rescaling  $\epsilon$  yields the required bound.

Time and space can be bound the same way as in [20] by keeping an orthonormal basis of  $\mathcal{S}A$ . However, note that  $\mathcal{S}$  is independent of the input and hence we can multiply  $\mathcal{S}$  with  $A$  in the first pass, compute  $\Pi_{\mathcal{S}A}(A)$  in the second and obtain  $\Pi_{\mathcal{S}A,k}(A)$  in two passes altogether.  $\square$ .

**Remark.** Since  $\Pi_{\mathcal{S}A,k}(A)$  is indeed computed as a sequence of two projections it is easy to keep track of the error using  $\|A\|_F^2 = \|A - \Pi_{\mathcal{S}A,k}(A)\|_F^2 + \|\Pi_{\mathcal{S}A,k}(A)\|_F^2$ . Thus we can boost the probability of success to  $1 - \delta$  by running  $O(\log(1/\delta))$  independent copies parallel and choosing the instance with maximal  $\|\Pi_{\mathcal{S}A,k}(A)\|_F^2$ .

Moreover the number of random bits required to construct  $\mathcal{S}$  can be reduced by showing that only the entries within the first  $\Theta(k)$  rows of  $\mathcal{S}$  need to be completely independent and that the remaining  $\Theta(k/\epsilon)$  rows can also contain four-wise independent tug-of-war vectors. The essence of the proof deferred to Appendix A is that we analyze the effect of the aforementioned submatrices of  $\mathcal{S}$  separately by showing that the adaptive sampling theorem of Deshpande et al. [20] holds with tug-of-war projections as well and then apply Theorem 14 with  $\epsilon = 1$  only.

Deshpande and Vempala also proved [21] that for any matrix  $A$ , there exists a subset  $R$  of  $O(k \log k + k/\epsilon)$  rows of  $A$  such that  $\|A - \Pi_{R,k}(A)\|_F \leq (1 + \epsilon) \|A - A_k\|_F$  and their approximate SVD method indeed finds an  $O(k^2 \log k + k/\epsilon)$  element row set (see also [30]). Combining Claim 13 with Theorem 14 it follows that if we sample according to the squared row lengths of  $V_k$  then in  $O(SVD_k(A))$  time we can find an  $O(k \log k + k/\epsilon)$  element column set  $C = AS^T$  such that  $\|A - \pi_C(A)\|_F \leq \|A - \pi_{C,k}(A)\|_F \leq (1 + \epsilon) \|A - A_k\|_F$ . It is easy to see that Theorem 14 and hence the previous inequality holds unchanged if we replace  $A_k$  with any matrix  $B_k$  such that  $B_k = \pi_X(A)$ , where  $X$  is a  $k$ -dimensional subspace. Thus

we can obtain a faster relative-error column-based approximation algorithm by applying Theorem 14 twice and sampling according to the row lengths of  $V_{\Pi_{S,A,k}(A)}$  in time  $O((M(k \log k + k/\epsilon) + (n + m)(k \log k + k/\epsilon)^2) \log(1/\delta))$  and 4 passes altogether.

Lastly, by a result of Drineas and Mahoney [27] Theorem 14 also yields improved low-rank approximation of higher order tensors in the “unfolding” model.

## 5 Conclusion

We conclude with two open problems. Does there exist a fast, pass efficient algorithm for  $(1 + \epsilon)\sigma_{k+1}$  relative-error low-rank approximation in the spectral norm? What space and time lower bounds can be proven for *any* pass efficient approximate matrix product,  $\ell_2$  regression, or SVD algorithm? And lastly, from a practical point of view, it is imperative to evaluate and compare the algorithms discussed in this paper using large scale synthetic and real world data.

## Acknowledgements

We wish to thank András A. Benczúr and Katalin Friedl for many fruitful discussions and numerous suggestions for improving the presentation and Piotr Indyk for pointing out references [45, 52].

## References

- [1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.
- [2] D. Achlioptas, A. Fiat, A. R. Karlin, and F. McSherry. Web search via hub synthesis. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 500–509, 2001.
- [3] D. Achlioptas and F. McSherry. Fast computation of low-rank approximations. *To appear in the Journal of the ACM*, 2003.
- [4] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005.
- [5] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th ACM Symposium on Theory of Computing (STOC)*, 2006.
- [6] N. Alon, P. B. Gibbons, Y. Matias, and M. Szegedy. Tracking join and self-join sizes in limited storage. *Journal of Computer and System Sciences*, 64(3):719–747, 2002.
- [7] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.
- [8] S. Arora, E. Hazan, and S. Kale. A fast random sampling algorithm for sparsifying matrices. In *Proceedings of the 10th International Workshop on Randomization and Computation (RANDOM)*, 2006.
- [9] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 616–623, 1999.
- [10] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of the 33rd ACM Symposium on Theory of Computing (STOC)*, pages 619–626, 2001.
- [11] Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proceedings of the 35th ACM Symposium on Theory of Computing (STOC)*, pages 335–344, 2003.
- [12] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. The Johnson-Lindenstrauss lemma meets compressed sensing, 2006. Preprint.
- [13] M. W. Berry, S. T. Dumais, and G. W. O’Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.
- [14] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 693–703, 2002.

- [15] E. Cohen and D. D. Lewis. Approximating matrix multiplication for pattern recognition tasks. *Journal of Algorithms*, 30(2):211–252, 1999.
- [16] H. Cohn, R. Kleinberg, B. Szegedy, and C. Umans. Group-theoretic algorithms for matrix multiplication. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 379–388, 2005.
- [17] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9(3):251–280, 1990.
- [18] A. Dasgupta, R. Kumar, P. Raghavan, and A. Tomkins. Variable latent semantic indexing. In *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 839–842, 2005.
- [19] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.
- [20] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1117–1126, 2006.
- [21] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Proceedings of the 10th International Workshop on Randomization and Computation (RANDOM)*, 2006.
- [22] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, pages 9–33, 2004.
- [23] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36:132–157, 2006.
- [24] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36:184–206, 2006.
- [25] P. Drineas, I. Kerenidis, and P. Raghavan. Competitive recommendation systems. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC)*, pages 82–90, 2002.
- [26] P. Drineas and M. W. Mahoney. Approximating a Gram matrix for improved kernel-based learning. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 323–337, 2005.
- [27] P. Drineas and M. W. Mahoney. A randomized algorithm for a tensor-based generalization of the SVD, 2005. To appear in *Linear Algebra and Its Applications*.
- [28] P. Drineas, M. W. Mahoney, and R. Kannan. Fast Monte Carlo algorithms for matrices II: Computing a low rank approximation to a matrix. *SIAM Journal on Computing*, 36:158–183, 2006.
- [29] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for  $\ell_2$  regression and applications. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1127–1136, 2006.
- [30] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *Proceedings of the 10th International Workshop on Randomization and Computation (RANDOM)*, 2006.
- [31] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *Proceedings of the 14th Annual European Symposium on Algorithms (ESA)*, 2006.
- [32] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures and Algorithms*, 27(2):251–275, 2005.
- [33] P. Frankl and H. Maehara. The Johnson-Lindenstrauss Lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B*, 44:355–362, 1988.
- [34] R. Freivalds. Probabilistic machines can use less running time. In *Proceedings of the IFIP Congress 1977*, pages 839–842, 1977.
- [35] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low rank approximations. In *Proceedings of the 39th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 370–378, 1998.
- [36] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1983.
- [37] S. Har-Peled. Low rank matrix approximation in linear time, 2006. Manuscript.
- [38] M. R. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. In *External Memory Algorithms, DIMACS Book Series vol. 50.*, pages 107–118. American Mathematical Society, 1999.
- [39] P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of Algorithms*, 53(3):307–323, 2006.
- [40] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

- [41] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [42] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 444–457, 2005.
- [43] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [44] J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl.*, 13(4):1094–1122, 1992.
- [45] A. Magen. Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications. In *Proceedings of the 6th International Workshop on Randomization and Computation (RANDOM)*, pages 239–253, 2002.
- [46] P.-G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the approximation of matrices. Technical Report 1361, Yale University, 2006.
- [47] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 529–537, 2001.
- [48] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.
- [49] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis, 2005. Submitted.
- [50] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 297–306, 2006. Full version available at <http://www.ilab.sztaki.hu/websearch/Publications/>.
- [51] S. J. Szarek. Spaces with large distance to  $\ell_\infty^n$  and random matrices. *American Journal of Mathematics*, 112(6):899–942, 1990.
- [52] H. Wasserman and M. Blum. Software reliability via run-time result-checking. *Journal of the ACM*, 44(6):826–849, 1997.

## Appendix

### A Relative-error SVD with fewer random bits

In this section we reduce the number of random bits required in Theorem 14 from  $O(\frac{k}{\epsilon}m')$  to  $O(km' + \frac{k}{\epsilon} \log m')$ , where  $m' = \min\{n, m\}$ , by proving Theorem 17. We remark that it is conceivable that the number of random bits required might be further reduced by using the results of [39]. First we recall one of the key statements of [20]. In what follows let  $\text{span}(A)$  denote  $\text{rowspan}(A)$  for short.

**Theorem 15 (Adaptive sampling, Theorem 6 in [20])** *Let  $A \in \mathbb{R}^{m \times n}$  and  $V \leq \mathbb{R}^n$  be an arbitrary vector subspace. Define  $E = A - \Pi_V(A)$ . Let  $S$  be a random sample of  $s$  rows from  $A$  such that row  $i$  is chosen with probability  $p_i = \frac{\|E_{(i)}\|_2^2}{\|E\|_F^2}$ . Then, for any nonnegative integer  $k$ ,*

$$\mathbf{E} \left( \|A - \Pi_{V \cup \text{span}(S), k}(A)\|_F^2 \right) \leq \|A - A_k\|_F^2 + \frac{k}{s} \|E\|_F^2.$$

After submitting the first version of this paper we learned about the independent results of Har-Peled [37] and Deshpande and Vempala [21] on relative-error low-rank matrix approximation. Conceptually [37] and [21] work as follows. First, through multiple pass sampling a subspace  $V_0$  is found such that  $\|E_0\|_F^2 = \|A - \Pi_{V_0}(A)\|_F^2$  is at most  $f(k) \cdot \|A - A_k\|_F^2$  for some function  $f$ . In [21] further rounds of sampling extend  $V_0$  to  $V_1$  such that  $\|E_1\|_F^2 = \|A - \Pi_{V_1}(A)\|_F^2 \leq O(1) \cdot \|A - A_k\|_F^2$ . Finally Theorem 15 is applied with the subspace  $V_1$  and  $O(k/\epsilon)$  samples reducing the error to  $(1 + \epsilon) \|A - A_k\|_F$ . Har-Peled [37] improves  $f(k)$  to  $1 + \epsilon$  directly using a technique derived from Theorem 15. The key contributions of [37, 21] are finding the large relative-error  $V_0$  subspaces.

We sharpen our analysis by using Theorem 15 as well. However, via Theorem 14 we find the  $O(1)$ -factor approximation subspace  $V_1 = \text{span}(S_1 A) \leq \mathbb{R}^n$  more efficiently in 2 passes by projecting the columns of the input matrix  $A$  to  $O(k)$  dimensions using an  $S_1 \in \mathbb{R}^{O(k) \times m}$  Johnson-Lindenstrauss matrix.

Next, we could immediately apply Theorem 15 and obtain a hybrid random projection and sampling based method that runs in time  $O((Mk/\epsilon + (n + m)k^2/\epsilon^2) \log(1/\delta))$  and 4 passes. Instead, by adapting Theorem 15 to random projections and showing that a single projection is just as powerful as multiple rounds of adaptive sampling, we derive a 2 pass algorithm with the same running time.

**Claim 16** *Let  $A \in \mathbb{R}^{m \times n}$  and  $V \leq \mathbb{R}^n$  be an arbitrary vector subspace. Define  $E = A - \Pi_V(A)$ . Let  $0 < \epsilon \leq 1$  and  $\mathcal{S} \in \mathbb{R}^{\epsilon^{-1} \times m}$  be a tug-of-war matrix. Then, for any nonnegative integer  $k$ ,*

$$\mathbf{E} \left( \|A - \Pi_{V \cup \text{span}(\mathcal{S}E), k}(A)\|_F^2 \right) \leq \|A - A_k\|_F^2 + 2k\epsilon \|E\|_F^2.$$

**PROOF:** We rephrase the essence of the proofs of [35, 20] in a compact manner that applies to sampling and projections as well. Set  $D = U_k U_k^T \Pi_V(A) + U_k U_k^T \mathcal{S}^T \mathcal{S} E$ . Similarly to the proof of Theorem 14 observe that  $\text{rank}(D) \leq k$  and that the rows of  $D$  lie in  $V \cup \text{span}(\mathcal{S}E)$ . Therefore

$$\|A - \Pi_{V \cup \text{span}(\mathcal{S}E), k}(A)\|_F^2 \leq \|A - D\|_F^2 = \|A - A_k\|_F^2 + \|A_k - D\|_F^2, \quad (9)$$

where the last equality follows from the fact that the columns of  $D$  lie in the column space of  $U_k$ . The unitary invariance of the Frobenius norm and  $A_k = U_k U_k^T A$  gives us

$$\|A_k - D\|_F^2 = \|U_k^T A - U_k^T \Pi_V(A) - U_k^T \mathcal{S}^T \mathcal{S} E\|_F^2 = \|U_k^T E - U_k^T \mathcal{S}^T \mathcal{S} E\|_F^2. \quad (10)$$

From the second statement of Lemma 6 and  $\|U_k\|_F^2 = k$  it follows that  $\mathbf{E} \left( \|U_k^T E - U_k^T \mathcal{S}^T \mathcal{S} E\|_F^2 \right) \leq 2\epsilon k \|E\|_F^2$ . Combining the latter with equations (9–10) concludes the proof. For the original, sampling based version of the theorem, see Theorem 15 and [20], we apply the column-based variant of Lemma 19 instead of Lemma 6 in the last step.  $\square$

**Theorem 17** *Let  $A \in \mathbb{R}^{m \times n}$  and  $0 < \epsilon \leq 1$  and  $\mathcal{S} \in \mathbb{R}^{O(k/\epsilon) \times m}$  be a tug-of-war matrix such that the entries of the first  $O(k)$  rows are completely independent forming a  $O(k) \times m$  Johnson-Lindenstrauss submatrix. Then with probability at least  $1/4$  it holds that*

$$\|A - \Pi_{\text{span}(\mathcal{S}A), k}(A)\|_F^2 \leq (1 + \epsilon) \|A - A_k\|_F^2.$$

PROOF: Let  $\mathcal{S} = \begin{bmatrix} \mathcal{S}_1 \\ \mathcal{S}_2 \end{bmatrix}$  with  $\mathcal{S}_1 \in \mathbb{R}^{O(k) \times m}$  and  $\mathcal{S}_2 \in \mathbb{R}^{O(k/\epsilon) \times m}$ . Define  $V_1 = \text{span}(\mathcal{S}_1 A)$  and  $V_2 = \text{span}(\mathcal{S}_2(A - \Pi_{V_1}(A)))$ . As  $\Pi_{V_1}(A) = X(\mathcal{S}_1 A)$  and  $\Pi_{V_2}(A) = Y(\mathcal{S}_2(A - X\mathcal{S}_1 A))$  for some  $X, Y$ , for some  $W, Z$  we have that

$$\Pi_{\text{span}(V_1 \cup V_2)}(A) = W\Pi_{\text{span}(V_1)}(A) + Z\Pi_{\text{span}(V_2)}(A) = WX(\mathcal{S}_1 A) + ZY(\mathcal{S}_2 A) - ZYX(\mathcal{S}_1 A) \in \text{span}(\mathcal{S}_1 A \cup \mathcal{S}_2 A).$$

Thus  $\|A - \Pi_{\text{span}(\mathcal{S}A), k}(A)\|_F \leq \|A - \Pi_{\text{span}(V_1 \cup V_2), k}(A)\|_F$ . From Theorem 14 applied to  $\mathcal{S}_1$  with  $\epsilon' = 1$  combined with  $\|A - \Pi_{V_1}(A)\|_F \leq \|A - \Pi_{V_1, k}(A)\|_F$  we have that

$$\Pr(\|A - \Pi_{V_1}(A)\|_F \leq 2\|A - A_k\|_F) \geq 1/2.$$

Conditioning on this event Claim 16 with  $V_1$  and  $\mathcal{S}_2$  and  $\epsilon'$  set to  $\epsilon/(16k)$  gives us

$$\mathbf{E} \left( \|A - \Pi_{\text{span}(\mathcal{S}A), k}(A)\|_F^2 - \|A - A_k\|_F^2 \right) \leq \frac{\epsilon}{2} \|A - A_k\|_F^2.$$

We conclude the proof by applying Markov's inequality to  $\|A - \Pi_{\text{span}(\mathcal{S}A)}(A)\|_F^2 - \|A - A_k\|_F^2$  and observing that  $\|A - \Pi_{\text{span}(\mathcal{S}A), k}(A)\|_F \leq \sqrt{(1 + 2\epsilon) \|A - A_k\|_F^2} \leq (1 + \epsilon) \|A - A_k\|_F$  holds with probability at least  $1/4$ .  $\square$

## B Further discussion of approximate matrix products

In this section we prove Theorem 9 and provide further discussion on approximating multiple term matrix products.

**Proof of Theorem 9** First we observe that picking the minimum of a few good enough approximations is close to the real minimum.

**Fact 18** *Let  $0 < \delta < 1$ ,  $0 < \lambda \leq 1/2$  and  $Y_i$ ,  $i = 1 \dots t$ , be random variables with expectation  $\mu_i$  and assume that  $\Pr(|Y_i - \mu_i| > \lambda\mu_i) \leq \delta/t$  holds. Let  $Y_{i^*} = \min_i Y_i$  and  $\mu_{i^{**}} = \min_i \mu_i$ . Then  $\Pr(\mu_{i^*} \leq \mu_{i^{**}}(1 + 4\lambda)) \geq 1 - \delta$ .*

PROOF: By the union bound  $\Pr(\forall i : |Y_i - \mu_i| \leq \lambda\mu_i) \geq 1 - \delta$ . In that case  $\mu_{i^*}(1 - \lambda) \leq Y_{i^*} \leq Y_{i^{**}} \leq \mu_{i^{**}}(1 + \lambda)$  holds. Thus  $\mu_{i^*} \leq \mu_{i^{**}} \frac{1+\lambda}{1-\lambda} \leq \mu_{i^{**}}(1 + 4\lambda)$ .  $\square$

PROOF:[Theorem 9] By Lemma 6 the algorithm is unbiased for all  $i$ . From Lemma 8 with  $C = AB - AS_i^T \mathcal{S}_i B$  it follows that  $\mathbf{E}(y_{i,j}) = \|AB - AS_i^T \mathcal{S}_i B\|_F^2$  and  $\mathbf{Var}(y_{i,j}) \leq 2(1/4^2) \|AB - AS_i^T \mathcal{S}_i B\|_F^4$ . By Chebyshev's inequality

$$\Pr\left(|y_{i,j} - \|AB - AS_i^T \mathcal{S}_i B\|_F^2| \geq \sqrt{2}\sqrt{2}(1/4) \|AB - AS_i^T \mathcal{S}_i B\|_F^2\right) \leq (1/\sqrt{2})^2.$$

Thus

$$\Pr\left(|z_i - \|AB - AS_i^T \mathcal{S}_i B\|_F^2| \geq 1/2 \|AB - AS_i^T \mathcal{S}_i B\|_F^2\right) \leq \frac{\delta}{\log 1/\delta}.$$

Let  $i^{**} = \operatorname{argmin}_i \|AB - AS_i^T \mathcal{S}_i B\|_F^2$ , by Lemma 18

$$\Pr\left(\|AB - AS_{i^*}^T \mathcal{S}_{i^*} B\|_F^2 \leq \|AB - AS_{i^{**}}^T \mathcal{S}_{i^{**}} B\|_F^2 (1 + 4(1/2))\right) \geq 1 - \delta.$$

Recall that by Fact 7 we have that  $\Pr\left(\|AB - AS_{i^{**}}^T \mathcal{S}_{i^{**}} B\|_F^2 \leq 4\epsilon^2 \|A\|_F^2 \|B\|_F^2\right) \geq 1 - \delta$ , and hence by the union bound

$$\Pr\left(\|AB - AS_{i^*}^T \mathcal{S}_{i^*} B\|_F^2 \leq 3 \cdot 4\epsilon^2 \|A\|_F^2 \|B\|_F^2\right) \geq 1 - 2\delta. \quad \square$$

**Remark.** If it is guaranteed that the entries of  $B$  precede the entries of  $A$ , then we can run the unmodified algorithm in a one pass streaming fashion. Without the  $A$  follows  $\widehat{B}$  assumption, it requires two passes.

It is easy to see that if we replace  $AS_i^T$  with  $\widehat{A}_i$  and  $\mathcal{S}_i B$  with  $\widehat{B}_i$  in Algorithm 1, where  $\widehat{A}_i$  and  $\widehat{B}_i$  are the output of the  $i$ th independent instance of *any* algorithm with  $\mathbf{E}\left(\|AB - \widehat{A}\widehat{B}\|_F^2\right)$  bounded then the good approximation with high probability part of Theorem 9 holds. It is also straightforward to modify Algorithm 1 to boost the success probability of any algorithm approximating  $p(A_1, \dots, A_u)$  with  $q(B_1, \dots, B_v)$ , where  $A_i, B_j$  are arbitrary matrices and  $p, q$  are polynomials. For example, returning to matrix products, we can also plug the following in.

**Lemma 19 (Lemma 8 in appendix A.2 of [23])** *Let the probability of picking column-row pair  $k$  be  $p_k = \frac{\|A^{(k)}\|_2^2}{\|A\|_F^2}$ . Pick  $1/\epsilon^2$  column-row pairs and form matrices  $C, R$  as in BasicMatrixMultiplication of [23]. Then  $\mathbf{E}\left(\|AB - CR\|_F^2\right) \leq \epsilon^2 \|A\|_F^2 \|B\|_F^2$ .*

## B.1 Product of more than two matrices

In this section we observe that Algorithm 1 in itself and also with Lemma 19 can be used unaltered to approximate multiple term matrix products as well. In contrast, one cannot apply column-row sampling [23] and use  $ABC = A(BC)$  directly to approximate  $BC$  first and then  $ABC$  since we need to sample non uniformly according to the row lengths from  $\widehat{B}\widehat{C}$ , which is not clear how to do, if the elements of  $\widehat{B}\widehat{C}$  are not available (computing the product takes quadratic space and time). Appendix A.1 of [23] suggests an extension of sampling for the product of three matrices, however it concludes that the computation of the sampling probabilities requires quadratic space and time in general. Nevertheless in Appendix B.2 we show

that it is possible to reduce  $k$  term matrix product approximation to sampling two term products at a price of  $\Omega(k^2)$  blowup in space. The resulting procedure is more cumbersome than that of this section and has weaker error bounds.

Using  $A_1 A_2 \dots A_l = (A_1 \dots A_{k-1}) A_k$  Algorithm 1 requires one pass if the matrix elements arrive in  $A_k, \dots, A_1$  order and comes with the following bound

$$\Pr \left( \|A_1 \dots A_k - A_1 \dots A_{k-1} \mathcal{S}_{i^*}^T \mathcal{S}_{i^*} A_k\|_F \leq \sqrt{12} \epsilon \|A_1 \dots A_{k-1}\|_F \|A_k\|_F \right) \geq 1 - 2\delta.$$

Or it finishes in  $k$  passes independent of the element order and we can break  $A_1 \dots A_k$  at an arbitrary position.

For Algorithm 1 combined with Lemma 19 observe that  $A_1 A_2 \dots A_k = A_1 (A_2 \dots A_k)$  and then we can select the  $i$ th row of  $(A_2 \dots A_k)$  by computing the product  $e_i^T A_2 \dots A_k$ . Hence it requires one pass if  $A_1$  comes first column grouped and the remaining matrix elements arrive in  $A_2, \dots, A_k$  order and comes with the following bound

$$\Pr \left( \|A_1 \dots A_k - C_{i^*} R_{i^*}\|_F \leq \sqrt{6} \epsilon \|A_1\|_F \|A_2 \dots A_k\|_F \right) \geq 1 - 2\delta.$$

Or it finishes in  $k$  passes independent of the element order.

## B.2 Approximating multiple term matrix products with column-row sampling

In this section we give an extension of the column-row sampling technique of Drineas et al. [23] for approximating matrix chain products. First we generalize Lemma 6 from [26] for approximating even products and observe that time and space complexity grow quadratically with the number of matrices involved. Then we reduce odd chains to even ones with a rather costly SVD computation. Additionally, in both cases the obtained bounds are of weaker form than that of Section B.1 and hence we present these result primarily for comparison.

To approximate even  $A_1 B_1 \dots A_l B_l$  products let us sample  $A_i$  and  $B_i$  (obtaining  $C_i, R_i$ ) as if we were to compute  $A_i B_i$  and then write  $C_1 R_1 C_2 R_2 \dots C_l R_l$  as  $C_1 \cdot ((R_1 C_2)(R_2 C_3) \dots (R_{l-1} C_l) R_l)$ .

**Lemma 20** *If we sample  $\Omega(\epsilon^{-2} \log(1/\delta))$  column-row pairs for each term  $A_i B_i$  then*

$$\Pr \left( \|A_1 B_1 \dots A_l B_l - C_1 R_1 \dots C_l R_l\|_F \leq ((1 + \epsilon)^l - 1) \prod_{i=1}^l \|A_i\|_F \|B_i\|_F \right) \geq 1 - l\delta.$$

PROOF: By the union bound for all  $i$   $\|C_i R_i\|_F \leq \|A_i B_i\|_F + \|C_i R_i - A_i B_i\|_F \leq (1 + \epsilon) \|A_i\|_F \|B_i\|_F$ . Note that  $A_1 B_1 \dots A_l B_l - C_1 R_1 \dots C_l R_l = \sum_{i=1}^l (\prod_{j=1}^{i-1} C_j R_j) (A_i B_i - C_i R_i) (\prod_{k=i+1}^l C_k R_k)$ . Hence

$$\|A_1 B_1 \dots A_l B_l - C_1 R_1 \dots C_l R_l\|_F \leq \left( \prod_{i=1}^l \|A_i\|_F \|B_i\|_F \right) \epsilon \sum_{i=1}^l (1 + \epsilon)^{i-1} = ((1 + \epsilon)^l - 1) \prod_{i=1}^l \|A_i\|_F \|B_i\|_F. \quad \square$$

For  $l$  not too large  $(1 + \epsilon)^l \approx 1 + l\epsilon$ . Thus we need to run column sampling with parameters  $(\epsilon/l, \delta/l)$ , which results in a factor  $l^2$  blowup in space. If  $A_i \in \mathbb{R}^{m_i \times n_i}$  and  $B_i \in \mathbb{R}^{n_i \times m_{i+1}}$  then the time of the "cheap"  $R_i C_{i+1}$ ,  $(l^2/\epsilon^2 \times n_i) \cdot (n_i \times l^2/\epsilon^2)$  term is  $O(l^4/\epsilon^4 n_i)$ , assuming that we compute matrix products naively. This corresponds to a factor  $l^4/\epsilon^2$  increase with respect to the time needed just to sample these column-row

pairs. Unless  $A_i = B_i^T$ , when  $\|C_i\|_F = \|R_i\|_F = \|A_i\|_F$ , no bound is known for  $\|C_i\|_F \cdot \|R_i\|_F$  in terms of  $\|A_i\|_F \cdot \|B_i\|_F$  and thus we cannot go further and sample the  $R_i C_{i+1}$  products as well. In any case, by submultiplicity of the matrix norm the bound obtained via sampling is weaker than that of Section B.1.

Turning to odd chains observe that inserting an extra  $n \times n$  identity matrix increases error by a  $\sqrt{n}$  factor. Hence we rather reduce length  $k = 2l + 1$  odd chains to even products by running an approximate SVD in the spectral norm [28, 3] on  $A_1 \in \mathbb{R}^{n_1 \times n_2}$  to obtain matrices  $X \in \mathbb{R}^{n_1 \times r}$ ,  $Y \in \mathbb{R}^{r \times n_2}$ , such that  $\Pr(\|A_1 - XY\|_2 \leq \|A_1 - (A_1)_r\|_2 + \epsilon \|A_1\|_F) \geq 1 - \delta/2$ . Note that

$$\|A_1 - (A_1)_r\|_2^2 = \sigma_{r+1}^2(A_1) \leq \sigma_r^2(A_1) \leq \sum_{i=1}^r \sigma_i^2(A_1)/r \leq \sum_{i=1}^{\text{rank}(A_1)} \sigma_i^2(A_1)/r = \|A_1\|_F^2 / r.$$

Hence for any  $0 < \epsilon \leq 1$  with  $r = 1/\epsilon^2$ ,  $\|A_1 - XY\|_2 \leq 2\epsilon \|A_1\|_F$ . Let  $B = A_2 \dots A_{2l+1}$ , and run the approximate matrix multiplication algorithm on this even product, and obtain  $C, R$  such that  $\Pr(\|B - CR\|_F \leq \epsilon \prod_{i=2}^{2l+1} \|A_i\|_F) \geq 1 - \delta/2$ . Finally we approximate  $\prod_{i=1}^{2l+1} A_i = A_1 B$  as  $X \cdot ((YC)R)$ , where all but the last multiplications are cheap.

Now let  $\Delta_{A_1} = A_1 - XY$  and  $\Delta_B = B - CR$  and note that by the union bound with probability at least  $1 - \delta$ ,  $\|\Delta_{A_1}\|_2 \leq 2\epsilon \|A_1\|_F$  and  $\|\Delta_B\|_F \leq \epsilon \prod_{i=2}^{2l+1} \|A_i\|_F$ . Additionally observe that for arbitrary matrices  $G, H$  it holds that  $\|GH\|_F \leq \|G\|_2 \|H\|_F$ . Hence

$$\begin{aligned} \|A_1 B - XYCR\|_F &= \|A_1 B - (A_1 + \Delta_{A_1})(B + \Delta_B)\|_F \\ &\leq \|\Delta_{A_1} B\|_F + \|A_1 \Delta_B\|_F + \|\Delta_{A_1} \Delta_B\|_F \\ &\leq \|\Delta_{A_1}\|_2 \|B\|_F + \|A_1\|_F \|\Delta_B\|_F + \|\Delta_{A_1}\|_2 \|\Delta_B\|_F \\ &\leq (2\epsilon + \epsilon + 2\epsilon^2) \prod_{i=1}^{2l+1} \|A_i\|_F \\ &\leq 5\epsilon \prod_{i=1}^{2l+1} \|A_i\|_F. \end{aligned}$$

To achieve the same error bound as for even chains we need to approximate the SVD with parameters  $(1/\epsilon^2, \epsilon, \delta/2)$  and the matrix multiplication with  $(\approx \epsilon/l, \delta/(2l))$ . Using either LinearTimeSVD or Constant-TimeSVD [28] for computing matrices  $X$  and  $Y$  requires  $O(n_1/\epsilon^4)$  additional space and  $O(n_1/\epsilon^8 + 1/\epsilon^{12})$  time.