

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 (ℓ_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*

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

- *The first $o(nd^2)$ time $(1 + \epsilon)$ relative error approximation algorithm for $n \times d$ linear (ℓ_2) regression.*
- *A matrix multiplication and norm approximation algorithm that easily applies to implicitly given matrices and can be used as a black box probability boosting tool.*

1. Introduction

This paper develops and analyzes fast approximation algorithms for fundamental linear algebra problems such

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

as singular value decomposition (SVD), linear ℓ_2 regression and the computation of matrix products. Our motivation comes from the widespread use of these tools in data mining [9]. Prominent applications of low rank matrix approximation by SVD include recommendation systems [23], information retrieval via Latent Semantic Indexing [11, 43], Kleinberg’s celebrated HITS algorithm for web search [39, 2], clustering [20, 42], and learning mixtures of distributions [38, 4] just to name a few. Classification can be solved by regularized regression [27] and text database querying by matrix-vector products [13].

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 [33], both of which are prohibitively large even for moderate size datasets arising in current applications. Even for sparse data it is often the case that the input far exceeds the main memory and hence we generally restrict ourselves to the pass efficient „streaming” model of computation [35]. 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 [33] 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 [40, 34].

Recently a large number of results appeared that prove bounds for non-uniform sampling to speed up approximate matrix operations [32, 43, 3, 21, 26, 22, 24, 44, 18]. 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 re-

sults of Drineas et al. [27, 28, 29], Har-Peled [34], and Deshpande et al. [18, 19]. 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, [27, 28, 29] 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 [12, 45]. Our key techniques to improve previous algorithms for singular value decomposition, ℓ_2 regression and matrix multiplication are Johnson-Lindenstrauss type embeddings [37]. Ironically, one of the first approximate singular value decomposition algorithms [43] was also embedding-based.

Our central result is a relative error SVD algorithm (Theorem 14). Extending the work of [28, 32, 18] we show that if we form $(k/\epsilon + k \log k)$ random linear combinations from the columns of $A \in \mathbb{R}^{m \times n}$, then the best rank- k approximation within the column 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 $O((M(k/\epsilon + k \log k) + (n + m)(k/\epsilon + k \log k)^2) \log(1/\delta))$. Independently of our work Har-Peled [34], and Deshpande and Vempala [19] also proved similar results. However, our procedure is faster in terms k than the more efficient of those, [19] that necessitates $\Theta(k \log k)$ passes.

We also present the first $o(nd^2)$ time $(1 + \epsilon)$ -approximation algorithm for ℓ_2 regression with coefficient matrix $A \in \mathbb{R}^{n \times d}$, $n \geq d = \omega(\log n)$, by replacing sampling in [27] with embeddings (Theorem 12). We offer novel analysis with improved bounds compared to [27], lowering the required number of reduced dimensions for sketches for example to $O(d \log d/\epsilon)$ that matches to the enhanced bound of [28] 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 ϵ down to $\omega(d(\log d + \log^2 n)/n)$.

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 [21] (Theorem 9). Unlike [21] 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 ℓ_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 [31] (Lemma 8). 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 (ℓ_2) regression results. These results are used finally in Section 4 in our SVD algorithm.

1.1. Comparison with previous results

Except for [43, 41], 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 [13]. In their ground-breaking paper Frieze, Kannan, and Vempala [32] 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 [26, 22, 44, 18] and approximate matrix products [21]. Other line of research is based on random sparsification and quantization [3, 21].

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 [33, 15, 14].

Returning to SVD, the best preliminary result with respect to the Frobenius norm was derived by Deshpande and Vempala [19] 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 [18], 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 num-

ber 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 [34], also independent of this work. Besides running in $O(\log n)$ passes it is slower than the other two approaches as its running time 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. [27] 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 ϵ -approximation to the original least squares problem. The same authors go a step further in [28] 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 [21]. In fact uniform sampling is insufficient as [10] shows. In [27, 28, 29] 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 ℓ_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 ρ 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 diag-

onal $\Sigma_k \in \mathbb{R}^{k \times k}$ contains first k entries of Σ . 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 $\pi_{V,k}(A)$ denote the best rank- k approximation of A with its columns in V , i.e. $\pi_{V,k}(A) = (\pi_V(A))_k$. Additionally, given matrix B let $\text{colspan}(B) \in \mathbb{R}^m$ denote the subspace generated by its columns and we use the simplified notation $\pi_{B,k}(A)$ for $\pi_{\text{colspan}(B),k}(A)$. 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 [33].

Random projections. Johnson-Lindenstrauss's seminal paper [37] was followed by several variants and proofs of low-distortion embeddings [30, 36, 17]. 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 ϵ, δ, d , or JLT(ϵ, δ, 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 $k = \Omega(\frac{\log d}{\epsilon^2} f(\delta))$ for all $v \in V$ holds that $(1 - \epsilon) \|v\|_2^2 \leq \|Rv\|_2^2 \leq (1 + \epsilon) \|v\|_2^2$ with probability at least $1 - \delta$.

Theorem 2 (The Johnson-Lindenstrauss Lemma [17, 8]) Let $0 < \epsilon, \delta < 1$ and $S = \frac{1}{\sqrt{k}} R \in \mathbb{R}^{k \times n}$ 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 S is a JLT(ϵ, δ, d).

For practical applications the $N(0, 1)$ entries can be replaced by random ± 1 variables [1, 8]. 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-L. Transform [5]) Let $S = \frac{1}{\sqrt{kn}} PH_n D$, where D is an $n \times n$ diagonal matrix with entries being independent uniformly random ± 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 ϵ_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 $(1 - \epsilon) \|v\|_2^2 \leq \|Sv\|_2^2 \leq (1 + \epsilon) \|v\|_2^2$.
- For all $x \in \mathbb{R}^n$ computing Sx takes $O(n \log n + \epsilon^{-2} \log^2 N \log d)$ time.

Now let us consider the dot product $\langle Su, Sv \rangle$ for $u, v \in V$. By the parallelogram rule it is easy to see [8, 43] that if S distorts squared norms by factor of at most $1 \pm \epsilon$ and the set V contains unit length vectors only then $|\langle Su, Sv \rangle - \langle u, v \rangle| \leq \epsilon$. If $u = 0$ then trivially $\langle S0, Sv \rangle = \langle 0, v \rangle = 0$ for all v . If $u \neq 0$ and $v \neq 0$ then by linearity $\langle Su, Sv \rangle = \|u\|_2 \|v\|_2 \left\langle \mathcal{S}_{\frac{u}{\|u\|_2}}, \mathcal{S}_{\frac{v}{\|v\|_2}} \right\rangle$ and therefore we also have the following stronger corollary, to which we will often refer to.

Corollary 4 *If S is a JLT(ϵ, δ, 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 $\langle u, v \rangle - \epsilon \|u\|_2 \|v\|_2 \leq \langle Su, Sv \rangle \leq \langle u, v \rangle + \epsilon \|u\|_2 \|v\|_2$.*

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 $\mathbf{E}(\|AB - AS^T SB\|_F^2) \leq 2\epsilon^2 \|A\|_F^2 \|B\|_F^2$.

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

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

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 [21].

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 S_i . Computing $\|AB - AS_i^T S_i B\|_F^2$ 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 $AS_i^T S_i B$. Similarly to Freivalds' technique for checking matrix products our norm estimation method requires a few extra matrix-vector products only [31] and was motivated by Lemma 4 of [16].

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

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

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 $\mathbf{E}(\langle x, r_i \rangle \langle y, r_i \rangle)^2 \leq 3\|x\|_2^2 \|y\|_2^2$. Hence

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

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.

Theorem 9 For Algorithm 1 we have that $\Pr\left(\|AB - AS_{i^*}^T S_{i^*} B\|_F \leq \sqrt{12}\epsilon \|A\|_F \|B\|_F\right) \geq 1 - 2\delta$ and $\mathbf{E}(AB - AS_{i^*}^T S_{i^*} B) = \mathbf{0}$. 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 omitted. Comparing Algorithm 1 to column-row sampling [21] 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. We defer further discussion to the full version of this paper.

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 .
-

3. The ℓ_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. [27, 28] 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 ϵ -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 [27] 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 [27], 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 [28]. 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. For

easier comparison with [27] we state the input parameters $\tilde{\epsilon}$ and \tilde{d} of the (F)JLT implicitly as $r = \Omega(\tilde{\epsilon}^{-2} \cdot \log \tilde{d})$ in the next lemma and theorem.

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). As a consequence we prove that given a k dimensional subspace V , embedding it into $O(k \log(k/\epsilon)/\epsilon^2)$ dimensions preserves the length of all vectors from V . 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 *Let V be an arbitrary k dimensional subspace of \mathbb{R}^m and $0 < \epsilon, \delta < 1$.*

If \mathcal{S} is a Johnson-Lindenstrauss transform from \mathbb{R}^m to $O(k \log(k/\epsilon)\epsilon^{-2} \cdot f(\delta))$ dimensions for some function f , then

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

PROOF: If $v = 0$ then trivially $\|\mathcal{S}0\|_2 = 0$ else by linearity $\|\mathcal{S}v\|_2 = \|v\|_2 \left\| \mathcal{S} \frac{v}{\|v\|_2} \right\|_2$ and hence w.l.o.g. we can assume that $\|v\|_2 = 1$. The key idea of the proof is that we „cover“ the unit sphere with a finite set H' such that each point of the unit sphere is close enough to an element of H' . So, let $\{u_i\}$ be an orthonormal basis of V and set $U = [u_1, \dots, u_k]$. Additionally let H be a $c = \min\{\sqrt{\epsilon}/k, \epsilon/\sqrt{k}\}$ -fine grid on $[-1, 1]^k$ and set $H' = \{Uh | h \in H\}$. Observe that $|H'| = O((k/\epsilon)^k)$. Applying Corollary 4 to $\{u_1, \dots, u_k\} \cup H'$, for arbitrary $Uy \in V$ we have that

$$\begin{aligned} \|\mathcal{S}Uy\|_2^2 &= \sum_{i=1}^k \sum_{j=1}^k \langle \mathcal{S}u_i y_i, \mathcal{S}u_j y_j \rangle \\ &\leq \|y\|_2^2 + \epsilon \sum_{i,j} |y_i| |y_j| = \|y\|_2^2 + \epsilon \|y\|_1^2. \end{aligned} \quad (1)$$

Note that $\|\mathcal{S}Uy\|_2$ is close to $\|y\|_2$ if $\|y\|_1$ is not too large. To exploit this observe that for any $x \in \mathbb{R}^k$ with $\|x\|_2 = 1$ there exists an $h \in H$ such that $\|h\|_2 \leq 1$, $\|x - h\|_2 \leq \sqrt{kc} \leq \epsilon$, and $\|x - h\|_1 \leq kc \leq \sqrt{\epsilon}$. It follows that $1 - \epsilon \leq \|h\|_2 = \|Uh\|_2$ and hence by the Johnson-Lindenstrauss property $1 - 2\epsilon \leq (1 - \epsilon)^2 \leq \|\mathcal{S}Uh\|_2 \leq 1 + \epsilon$.

Now, for an arbitrary point on the unit sphere $v = Ux \in V$, $\|x\|_2 = 1$, inequality (1) with $y = x - h$ gives us $\|\mathcal{S}U(x - h)\|_2^2 \leq \|x - h\|_2^2 + \epsilon \|x - h\|_1^2 \leq \epsilon^2 + \epsilon \sqrt{\epsilon}^2 = 2\epsilon^2$. Thus $\|\mathcal{S}Ux\|_2 \leq \|\mathcal{S}Uh\|_2 +$

$\|\mathcal{S}U(x - h)\|_2 \leq 1 + \epsilon + \sqrt{2}\epsilon \leq 1 + 3\epsilon$. Similarly $\|\mathcal{S}Ux\|_2 \geq \|\mathcal{S}Uh\|_2 - \|\mathcal{S}U(x - h)\|_2 \geq 1 - 4\epsilon$. Rescaling ϵ concludes the proof. \square

Corollary 11 *Let $0 < \epsilon, \delta < 1$ and \mathcal{S} be a Johnson-Lindenstrauss transform for some function f .*

- *If $U \in \mathbb{R}^{m \times k}$, $m \geq k$, is a unitary matrix and \mathcal{S} is a JLT from \mathbb{R}^m to $O(k \log(k/\epsilon)/\epsilon^2 \cdot f(\delta))$ dimensions, then*

$$\Pr(\forall i \in [1..k] : |1 - \sigma_i(\mathcal{S}U)| \leq \epsilon) \geq 1 - \delta.$$

- *(Weak) spectral bound for approximate matrix products. If $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$ and \mathcal{S} is a JLT from \mathbb{R}^n to $O(n \log(n/\epsilon)/\epsilon^2 \cdot f(\delta))$ dimensions, then*

$$\Pr(\|\mathcal{A}\mathcal{S}^T \mathcal{S}B - AB\|_2 \leq \epsilon \|A\|_2 \|B\|_2) \geq 1 - \delta.$$

PROOF: The first statement is just a reformulation of the fact that by Lemma 10 with high probability for any unit length vector $x \in \mathbb{R}^k$ it holds that $1 - \epsilon \leq \|\mathcal{S}Ux\|_2 \leq 1 + \epsilon$.

For the second statement observe that $\|\mathcal{A}\mathcal{S}^T \mathcal{S}B - AB\|_2 = \|A(\mathcal{S}^T \mathcal{S} - I_n)B\|_2 \leq \|A\|_2 \|B\|_2 \|\mathcal{S}^T \mathcal{S} - I_n\|_2$. As $\mathcal{S}^T \mathcal{S} - I_n$ is symmetric we have that $\|\mathcal{S}^T \mathcal{S} - I_n\|_2 = \max_{\|x\|_2=1} x^T (\mathcal{S}^T \mathcal{S} - I_n) x$ and hence it is sufficient to prove that $x^T \mathcal{S}^T \mathcal{S} x = \|\mathcal{S}x\|_2^2 \leq 1 + \epsilon$ holds with probability $1 - \delta$ for all unit length $x \in \mathbb{R}^n$. Applying Lemma 10 to $V = \mathbb{R}^n \leq \mathbb{R}^m$ with $\epsilon' = \epsilon/3$ establishes the claim. \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 $2/3$*

$$\tilde{\mathcal{Z}} \leq (1 + \epsilon)\mathcal{Z}. \quad (2)$$

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

$$\|b - A\tilde{x}_{opt}\|_2 \leq (1 + \epsilon)\mathcal{Z}. \quad (3)$$

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

$$\|x_{opt} - \tilde{x}_{opt}\|_2 \leq \frac{\epsilon}{\sigma_{\min}(A)} \mathcal{Z}. \quad (4)$$

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 (3) and $p = 2$ in the case of (4), thus for $d = \omega(\log n)$ we achieve (3) in $o(nd^2)$ time if $\epsilon = \omega(d(\log d + \log^2 n)/n)$.

PROOF:[Inequality 2] Applying the JLT to the single vector $b - Ax_{opt}$, by linearity we immediately obtain

$$\begin{aligned} \tilde{\mathcal{Z}} &= \|Sb - SA\tilde{x}_{opt}\|_2 \leq \|S(b - Ax_{opt})\|_2 \\ &\leq (1 + \epsilon) \|b - Ax_{opt}\|_2 = (1 + \epsilon)\mathcal{Z}. \end{aligned}$$

[Inequality 3] 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

$$\|b - A\tilde{x}_{opt}\|_2^2 = \|w - U\beta\|_2^2 = \mathcal{Z}^2 + \|\beta\|_2^2. \quad (5)$$

To upper bound $\|\beta\|_2^2$, recall that $\pi(\cdot)$ denotes the 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

$$U^T S^T \cdot SU\beta = U^T S^T \cdot Sw. \quad (6)$$

Now, as $r = \Omega(\rho \log \rho)$ setting ϵ' 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 (5) 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 4] 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

$$\begin{aligned} \sigma_{\min}(A) \|x_{opt} - \tilde{x}_{opt}\|_2 &= \sigma_{\min}(\Sigma V^T) \|\eta\|_2 = \\ \sqrt{\sum_{i=1}^{\rho} \sigma_i^2 \eta_i^2} &\leq \sqrt{\sum_{i=1}^{\rho} \sigma_i^2 \eta_i^2} = \|\beta\|_2 \leq 4\epsilon \mathcal{Z}. \quad \square \end{aligned}$$

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 (2-4) 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 (4), 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 (4) as the proof of inequality (3.16) in [27] shows.

We conclude this section by observing that the proof of inequalities (3) and (4) works unchanged for any matrix \mathcal{S} such that $|1 - \sigma_i^2(\mathcal{S}U)| = o(1)$ and $U\mathcal{S}^T Sw \approx U^T w$. Thus combining the above with Rudelson and Vershynin's proof of Theorem 1.1 in [44] for bounding the singular values and Lemma 8 in appendix A.2 of [21] for bounding the norm of the approximate matrix product we have the following claim for sampling ℓ_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 $\Pr \left(S_{(j)} = \frac{e_j}{\sqrt{r p_j}} \right) = p_j$ for all $1 \leq j \leq r$, where e_i denotes the i th unit vector. Then for any $0 < \epsilon \leq 1$ inequalities (3) and (4) 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 (2) 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 [27, 28]. The latter modification also yields $r = O(d \log(d)/\epsilon)$ for the sampling version of inequality (3), 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 [28] we show that if we form $(k/\epsilon + k \log k)$ random linear combinations of columns of A then the best rank- k approximation within the (column)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 $O((M(k/\epsilon + k \log k) + (n + m)(k/\epsilon + k \log k)^2) \log(1/\delta))$, 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 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 ± 1 entries and $r = \Theta(k/\epsilon + k \log k)$ then with probability at least $1/2$ it holds that*

$$\|A - \pi_{AS^T, k}(A)\|_F \leq (1 + \epsilon) \|A - A_k\|_F.$$

Computing the singular vectors spanning $\pi_{AS^T, k}(A)$ 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 [28]. Let $A = U\Sigma V^T$ be the SVD of A and $\rho = \text{rank}(A)$. Additionally set $X = \text{colspan}(\pi_{AS^T}(A_k))$ and denote its projector matrix by $P = (\pi_{AS^T}(A_k))(\pi_{AS^T}(A_k))^+$. Note that X is an at most k dimensional subspace in AS^T and since $\pi_{AS^T, k}(A)$ is the best rank- k approximation of A from $\text{colspan}(AS^T)$ we have

$$\|A - \pi_{AS^T, k}(A)\|_F^2 \leq \|A - PA\|_F^2.$$

By the unitary invariance of the Frobenius norm

$$\begin{aligned} \|A - PA\|_F^2 &= \|U\Sigma V^T - PU\Sigma V^T\|_F^2 \\ &= \|U\Sigma - PU\Sigma\|_F^2. \end{aligned}$$

By slightly abusing the notation and writing Σ as $\begin{bmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\rho-k} \end{bmatrix}$ it follows that

$$\begin{aligned} \|U\Sigma - PU\Sigma\|_F^2 &= \|U_k \Sigma_k - PU_k \Sigma_k\|_F^2 \\ &\quad + \|U_{\rho-k} \Sigma_{\rho-k} - PU_{\rho-k} \Sigma_{\rho-k}\|_F^2. \end{aligned}$$

As multiplying by a projection does not increase a unitarily invariant norm

$$\begin{aligned} \|U_{\rho-k} \Sigma_{\rho-k} - PU_{\rho-k} \Sigma_{\rho-k}\|_F^2 &= \\ \|(I - P)U_{\rho-k} \Sigma_{\rho-k}\|_F^2 &\leq \\ \|U_{\rho-k} \Sigma_{\rho-k}\|_F^2 &= \|A - A_k\|_F^2, \end{aligned}$$

since $I - P$ is projector matrix as well.

To complete the proof it is sufficient to show that $\|U_k \Sigma_k - PU_k \Sigma_k\|_F^2 \leq 2\epsilon \|A - A_k\|_F^2$ with probability at least $1/2$ since combining the latter with the previous four equations immediately gives us

$$\begin{aligned} \|A - \pi_{AS^T, k}(A)\|_F &\leq \sqrt{(1 + 2\epsilon) \|A - A_k\|_F^2} \\ &\leq (1 + \epsilon) \|A - A_k\|_F. \end{aligned}$$

For bounding $\|U_k \Sigma_k - PU_k \Sigma_k\|_F^2 = \|U_k \Sigma_k V_k^T - PU_k \Sigma_k V_k^T\|_F^2$ observe that $PU_k \Sigma_k V_k^T = PA_k = (AS^T)(AS^T)^+ A_k$ is the best approximation of A_k from the column space of AS^T and hence

$$\begin{aligned} \|A_k - PA_k\|_F^2 &\leq \|A_k - (AS^T)(A_k S^T)^+ A_k\|_F^2 \\ &= \|A_k^T - A_k^T (\mathcal{S} A_k^T)^+ (\mathcal{S} A^T)\|_F^2. \end{aligned} \quad (7)$$

Recall that $Y^{(i)}$ denotes the i th column of matrix Y and let us consider the regressions $A^{T(i)} \approx A_k^T x_i$ for $i = 1, \dots, m$. Note that the best approximation of $A^{T(i)}$ from A_k^T is $\pi_{A_k^T}(A^{T(i)}) = A_k^T (A^{T(i)})^+$ and hence it follows as equations (5-6) in the proof of Theorem 12 that there exists vectors $\beta_1, \dots, \beta_m \in \mathbb{R}^k$ and $w_1, \dots, w_m \in \mathbb{R}^n$ orthogonal to $\text{colspan}(V_k)$ such that

$$\begin{aligned} \forall i \in \{1, \dots, m\} : \|w_i\|_2^2 &= \|A^{T(i)} - A_k^T (A^{T(i)})^+\|_2^2 \\ \forall i \in \{1, \dots, m\} : V_k^T \mathcal{S}^T \mathcal{S} V_k \beta_i &= V_k^T \mathcal{S}^T \mathcal{S} w_i, \text{ and} \\ \sum_{i=1}^m \|\beta_i\|_2^2 &= \|A_k^T - A_k^T (\mathcal{S} A_k^T)^+ (\mathcal{S} A^T)\|_F^2 \end{aligned} \quad (8)$$

From $r = \Omega(k \log k)$ and Corollary 11 we have $\|\beta_i\|_2^2 / 2 \leq \|V_k^T \mathcal{S}^T \mathcal{S} V_k \beta_i\|_2^2$ with probability $3/4$ for all i as before. Observing that \mathcal{S} is a tug-of-war matrix as well and applying the second statement of Lemma 6

with $\epsilon' = \sqrt{\epsilon/d}$ to V_k^T and w_i from $V_k^T w_i = \mathbf{0}$ it follows that

$$\begin{aligned} \mathbf{E} \left(\sum_{i=1}^m \|V_k^T \mathcal{S}^T \mathcal{S} w_i\|_2^2 \right) &= \sum_{i=1}^m \mathbf{E} \left(\|V_k^T \mathcal{S}^T \mathcal{S} w_i\|_2^2 \right) \\ &\leq \sum_{i=1}^m \epsilon \|w_i\|_2^2 = \epsilon \sum_{i=1}^m \|A^{T^{(i)}} - A_k^{T^{(i)}}\|_2^2 \end{aligned}$$

Thus by Markov's inequality and the union bound we have that $\sum_{i=1}^m \|\beta_i\|_2^2 \leq 8 \|A^T - A_k^T\|_F^2$ holds with probability at least $1/2$. Combining the latter with equations (7-8) and rescaling ϵ yields the required bound.

Time and space can be bound the same way as in [18] by keeping an orthonormal basis of AS^T . However, note that \mathcal{S} is independent of the input and hence we can multiply A with \mathcal{S}^T in the first pass, compute $\pi_{AS^T}(A)$ in the second and obtain $\pi_{AS^T,k}(A)$ in two passes altogether. \square .

Remark. Since $\pi_{AS^T,k}(A)$ is indeed computed as a sequence of two projections it is easy to keep track of the error using $\|A\|_F^2 = \|A - \pi_{AS^T,k}(A)\|_F^2 + \|\pi_{AS^T,k}(A)\|_F^2$. 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 $\|\pi_{AS^T,k}(A)\|_F^2$.

Moreover the number of random bits required to construct \mathcal{S} can be reduced by showing that only the entries within the first $\Theta(k \log k)$ 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 omitted proof 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. [18] holds with tug-of-war projections as well and then apply Theorem 14 with $\epsilon = 1/2$ only.

Deshpande and Vempala also proved [19] 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 [28]). 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 $V_{\pi_{AS^T,k}(A)}$ in approximate SVD time and 4 passes altogether.

Lastly, by a result of Drineas and Mahoney [25] 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 $(1 + \epsilon)\sigma_k$ 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, ℓ_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.

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 *Proc. of the 42nd 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 *Proc. of the 18th COLT*, pages 458–469, 2005.
- [5] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proc. of the 38th 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] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proc. of the 40th FOCS*, pages 616–623, 1999.

- [9] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proc. of the 33rd STOC*, pages 619–626, 2001.
- [10] Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proc. of the 35th STOC*, pages 335–344, 2003.
- [11] 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.
- [12] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proc. of the 29th ICALP*, pages 693–703, 2002.
- [13] E. Cohen and D. D. Lewis. Approximating matrix multiplication for pattern recognition tasks. *Journal of Algorithms*, 30(2):211–252, 1999.
- [14] H. Cohn, R. Kleinberg, B. Szegedy, and C. Umans. Group-theoretic algorithms for matrix multiplication. In *Proc. of the 46th FOCS*, pages 379–388, 2005.
- [15] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9(3):251–280, 1990.
- [16] A. Dasgupta, R. Kumar, P. Raghavan, and A. Tomkins. Variable latent semantic indexing. In *Proc. of the 11th KDD*, pages 839–842, 2005.
- [17] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.
- [18] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proc. of the 17th SODA*, 2006.
- [19] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Proc. of the 10th RANDOM*, 2006.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] P. Drineas, I. Kerenidis, and P. Raghavan. Competitive recommendation systems. In *Proc. of the 34th STOC*, pages 82–90, 2002.
- [24] P. Drineas and M. W. Mahoney. Approximating a Gram matrix for improved kernel-based learning. In *Proc. of the 18th COLT*, pages 323–337, 2005.
- [25] 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*.
- [26] 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.
- [27] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proc. of the 17th SODA*, pages 1127–1136, 2006.
- [28] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *Proc. of the 10th RANDOM*, 2006.
- [29] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *Proc. of the 14th ESA*, 2006.
- [30] P. Frankl and H. Maehara. The Johnson-Lindenstrauss Lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B*, 44:355–362, 1988.
- [31] R. Freivalds. Probabilistic machines can use less running time. In *Proc. of the IFIP Congress 1977*, 1977.
- [32] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low rank approximations. In *Proc. of the 39th FOCS*, pages 370–378, 1998.
- [33] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1983.
- [34] S. Har-Peled. Low rank matrix approximation in linear time, 2006. Manuscript.
- [35] 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.
- [36] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. of the 30th STOC*, pages 604–613, 1998.
- [37] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [38] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proc. of the 18th COLT*, pages 444–457, 2005.
- [39] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [40] 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.
- [41] P.-G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the approximation of matrices. Technical Report 1361, Yale University, 2006.
- [42] F. McSherry. Spectral partitioning of random graphs. In *Proc. of the 42nd FOCS*, pages 529–537, 2001.
- [43] 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.
- [44] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis, 2005. Submitted.
- [45] 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 World Wide Web Conference (WWW)*, 2006.