

Low Rank Approximation and Regression in Input Sparsity Time

Kenneth L. Clarkson
IBM Almaden

David P. Woodruff
IBM Almaden

November 27, 2024

Abstract

We design a new distribution over $\text{poly}(r\varepsilon^{-1}) \times n$ matrices S so that for any fixed $n \times d$ matrix A of rank r , with probability at least $9/10$, $\|SAx\|_2 = (1 \pm \varepsilon)\|Ax\|_2$ simultaneously for all $x \in \mathbb{R}^d$. Such a matrix S is called a *subspace embedding*. Furthermore, SA can be computed in $O(\text{nnz}(A))$ time, where $\text{nnz}(A)$ is the number of non-zero entries of A . This improves over all previous subspace embeddings, which required at least $\Omega(nd \log d)$ time to achieve this property. We call our matrices S *sparse embedding matrices*.

Using our sparse embedding matrices, we obtain the fastest known algorithms for overconstrained least-squares regression, low-rank approximation, approximating all leverage scores, and ℓ_p -regression:

- to output an x' for which

$$\|Ax' - b\|_2 \leq (1 + \varepsilon) \min_x \|Ax - b\|_2$$

for an $n \times d$ matrix A and an $n \times 1$ column vector b , we obtain an algorithm running in $O(\text{nnz}(A) + \tilde{O}(d^3 \varepsilon^{-2}))$ time, and another in $O(\text{nnz}(A) \log(1/\varepsilon) + \tilde{O}(d^3 \log(1/\varepsilon)))$ time. (Here $\tilde{O}(f) = f \cdot \log^{O(1)}(f)$.)

- to obtain a decomposition of an $n \times n$ matrix A into a product of an $n \times k$ matrix L , a $k \times k$ diagonal matrix D , and an $n \times k$ matrix W , for which

$$\|A - LDW^\top\|_F \leq (1 + \varepsilon) \|A - A_k\|_F,$$

where A_k is the best rank- k approximation, our algorithm runs in

$$O(\text{nnz}(A)) + \tilde{O}(nk^2 \varepsilon^{-4} + k^3 \varepsilon^{-5})$$

time.

- to output an approximation to all leverage scores of an $n \times d$ input matrix A simultaneously, with constant relative error, our algorithms run in $O(\text{nnz}(A) \log n) + \tilde{O}(r^3)$ time.
- to output an x' for which

$$\|Ax' - b\|_p \leq (1 + \varepsilon) \min_x \|Ax - b\|_p$$

for an $n \times d$ matrix A and an $n \times 1$ column vector b , we obtain an algorithm running in $O(\text{nnz}(A) \log n) + \text{poly}(r\varepsilon^{-1})$ time, for any constant $1 \leq p < \infty$.

We optimize the polynomial factors in the above stated running times, and show various tradeoffs. Finally, we provide preliminary experimental results which suggest that our algorithms are of interest in practice.

1 Introduction

A large body of work has been devoted to the study of fast randomized approximation algorithms for problems in numerical linear algebra. Several well-studied problems in this area include least squares regression, low rank approximation, and approximate computation of leverage scores. These problems have many applications in data mining [5], recommendation systems [19], information retrieval [45], web search [1, 35], clustering [15, 38], and learning mixtures of distributions [34, 2]. The use of randomization and approximation allows one to solve these problems much faster than with deterministic methods.

For example, in the overconstrained least-squares regression problem, we are given an $n \times d$ matrix A of rank r as input, $n \gg d$, together with an $n \times 1$ column vector b . The goal is to output a vector x' so that with high probability, $\|Ax' - b\|_2 \leq (1 + \varepsilon) \min_x \|Ax - b\|_2$. The minimizing vector x^* can be expressed in terms of the Moore-Penrose pseudoinverse A^+ of A , namely, $x^* = A^+b$. If A has full column rank, this simplifies to $x^* = (A^\top A)^{-1}A^\top b$. This minimizer can be computed deterministically in $O(nd^2)$ time, but with randomization and approximation, this problem can be solved in $O(nd \log d) + \text{poly}(d\varepsilon^{-1})$ time [50, 26], which is much faster for $d \ll n$ and ε not too small. The generalization of this problem to ℓ_p -regression is to output a vector x' so that with high probability $\|Ax' - b\|_p \leq (1 + \varepsilon) \min_x \|Ax - b\|_p$. This can be solved exactly using convex programming, though with randomization and approximation it is possible to achieve $O(nd \log n) + \text{poly}(d\varepsilon^{-1})$ time [9] for any constant p , $1 \leq p < \infty$.

Another example is low rank approximation. Here we are given an $n \times n$ matrix (which can be generalized to $n \times d$) and an input parameter k , and the goal is to find an $n \times n$ matrix A' of rank at most k for which $\|A' - A\|_F \leq (1 + \varepsilon)\|A - A_k\|_F$, where for an $n \times n$ matrix B , $\|B\|_F^2 \equiv \sum_{i=1}^n \sum_{j=1}^n B_{i,j}^2$ is the squared Frobenius norm, and $A_k \equiv \text{argmin}_{\text{rank } B \leq k} \|A - B\|_F$. Here A_k can be computed deterministically using the singular value decomposition in $O(n^3)$ time. However, using randomization and approximation, this problem can be solved in $O(\text{nnz}(A) \cdot (k/\varepsilon + k \log k) + n \cdot \text{poly}(k/\varepsilon))$ time [50, 10], where $\text{nnz}(A)$ denotes the number of non-zero entries of A . The problem can also be solved using randomization and approximation in $O(n^2 \log n) + n \cdot \text{poly}(k/\varepsilon)$ time [50], which may be faster than the former for dense matrices and large k .

Another problem we consider is approximating the *leverage scores*. Given an $n \times d$ matrix A with $n \gg d$, one can write $A = U\Sigma V^\top$ in its singular value decomposition, where the columns of U are the left singular vectors, Σ is a diagonal matrix, and the columns of V are the right singular vectors. Although U has orthonormal columns, not much can be immediately said about the squared lengths $\|U_i\|_2^2$ of its rows. These values are known as the leverage scores, and measure the extent to which the singular vectors of A are correlated with the standard basis. The leverage scores are basis-independent, since they are equal to the diagonal elements of the projection matrix onto the span of the columns of A ; see [21] for background on leverage scores as well as a list of applications. The leverage scores will also play a crucial role in our work, as we shall see. The goal of approximating the leverage scores is to, simultaneously for each $i \in [n]$, output a constant factor approximation to $\|U_i\|_2^2$. Using randomization, this can be solved in $O(nd \log n + d^3 \log d \log n)$ time [21].

There are also solutions for these problems based on sampling. They either get a weaker additive error [27, 45, 3, 16, 17, 18, 22, 49, 13], or they get bounded relative error but are slow [14, 23, 24, 25]. Many of the latter algorithms were improved independently by Deshpande and Vempala [14] and Sarlós [50], and in followup work [26, 44, 37]. There are also solutions based on iterative and conjugate-gradient methods, see, e.g., [52], or [54] as recent examples. These methods repeatedly compute matrix-vector products Ax for various vectors x ; in the most common setting, such products require $\Theta(\text{nnz}(A))$ time. Thus the work per iteration of these methods is $\Theta(\text{nnz}(A))$, and the number of iterations N that are performed depends on the desired accuracy, spectral properties of A , numerical stability issues, and other concerns, and can be large. A recent survey suggests that N is typically $\Theta(k)$ for Krylov methods (such as Arnoldi and Lanczos iterations) to approximate the k leading singular vectors [29]. One can also use some of these techniques together, for example by first obtaining a preconditioner using the Johnson-Lindenstrauss (JL) transform, and then running an iterative method.

While these results illustrate the power of randomization and approximation, their main drawback is that they are not optimal. For example, for regression, ideally we could hope for $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ time. While the $O(nd \log d) + \text{poly}(d/\varepsilon)$ time algorithm for least squares regression is almost optimal for

dense matrices, if $\text{nnz}(A) \ll nd$, say $\text{nnz}(A) = O(n)$, as commonly occurs, this could be much worse than an $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ time algorithm. Similarly, for low rank approximation, the best known algorithms that are condition-independent run in $O(\text{nnz}(A)(k/\varepsilon + k \log k) + n \cdot \text{poly}(k/\varepsilon))$ time, while we could hope for $O(\text{nnz}(A)) + \text{poly}(k/\varepsilon)$ time.

1.1 Results

We resolve the above gaps by achieving algorithms for least squares regression, low rank approximation, and approximate leverage scores, whose time complexities have a leading order term that is $O(\text{nnz}(A))$, sometimes up to a log factor, with constant factors that are independent of any numerical properties of A . Our results are as follows:

- **Least Squares Regression:** We present several algorithms for an $n \times d$ matrix A with rank r and given $\varepsilon > 0$. One has running time bound of $O(\text{nnz}(A) \log(n/\varepsilon) + r^3 \log^2 r + r^2 \log(1/\varepsilon))$, stated at Theorem 43. (Note the logarithmic dependence on ε ; a variation of this algorithm has $O(\text{nnz}(A) \log(1/\varepsilon) + d^3 \log^2 d + d^2 \log(1/\varepsilon))$ running time.) Another has running time $O(\text{nnz}(A)) + \tilde{O}(d^3 \varepsilon^{-2})$, stated at Theorem 30; note that the dependence on $\text{nnz}(A)$ is linear. We also give an algorithm for generalized (multiple-response) regression, where $\min_X \|AX - B\|$ is found for $B \in \mathbb{R}^{n \times d'}$, in time

$$O(\text{nnz}(A) \log n + r^2((r + d')\varepsilon^{-1} + rd' + r \log^2 r + \log n));$$

see Theorem 38. We also note improved results for constrained regression, §7.6.

- **Low Rank Approximation:** We achieve running time $O(\text{nnz}(A)) + n \cdot \text{poly}(k(\log n)/\varepsilon)$ to find an orthonormal $L, W \in \mathbb{R}^{n \times k}$ and diagonal $D \in \mathbb{R}^{k \times k}$ matrix with $\|A - LDW^\top\|_F$ within $1 + \varepsilon$ of the error of the best rank- k approximation. More specifically, Theorem 47 gives a time bound of

$$O(\text{nnz}(A)) + \tilde{O}(nk^2\varepsilon^{-4} + k^3\varepsilon^{-5}).$$

- **Approximate Leverage Scores:** For any fixed constant $\varepsilon > 0$, we simultaneously $(1 + \varepsilon)$ -approximate all n leverage scores in $O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n)$ time. This can be generalized to sub-constant ε to achieve $O(\text{nnz}(A) \log n) + \text{poly}(r/\varepsilon)$ time, though in the applications we are aware of, such as coresets for regression [11], ε is typically constant (in the applications of this, a general $\varepsilon > 0$ can be achieved by over-sampling [23, 11]).
- **ℓ_p -Regression:** For $p \in [1, \infty)$ we achieve running time $O(\text{nnz}(A) \log n) + \text{poly}(r\varepsilon^{-1})$ in Theorem 50 as an immediate corollary of our results and a recent connection between ℓ_2 and ℓ_p -regression given in [9] (for $p = 2$, the $\text{nnz}(A) \log n$ term can be improved to $\text{nnz}(A)$ as stated above).

1.2 Techniques

All of our results are achieved by improving the time complexity of computing what is known as a *subspace embedding*. For a given $n \times d$ matrix A , call $S : \mathbb{R}^n \mapsto \mathbb{R}^t$ a *subspace embedding matrix* for A if, for all $x \in \mathbb{R}^d$, $\|SAx\|_2 = (1 \pm \varepsilon)\|Ax\|_2$. That is, S embeds the column space $C(A) \equiv \{Ax \mid x \in \mathbb{R}^d\}$ into \mathbb{R}^t while approximately preserving the norms of all vectors in that subspace.

The *subspace embedding problem* is to find such an embedding matrix obviously, that is, to design a distribution π over linear maps $S : \mathbb{R}^n \mapsto \mathbb{R}^t$ such that for any fixed $n \times d$ matrix A , if we choose $S \sim \pi$ then with large probability, S is an embedding matrix for A . The goal is to minimize t as a function of n, d , and ε , while also allowing the matrix-matrix product $S \cdot A$ to be computed quickly.

(A closely related construction, easily derived from a subspace embedding, is an *affine embedding*, involving an additional matrix $B \in \mathbb{R}^{n \times d'}$, such that

$$\|AX - B\|_F \approx \|S(AX - B)\|_F,$$

for all $X \in \mathbb{R}^{d \times d'}$; see §7.5. These affine embeddings are used for our low-rank approximation results, and immediately imply approximation algorithms for constrained regression.)

By taking S to be a Fast Johnson Lindenstrauss transform, one can set $t = O(d/\varepsilon^2)$ and achieve $O(nd \log t)$ time for $d < n^{1/2-\gamma}$ for any constant $\gamma > 0$. One can also take S to be a subsampled randomized Hadamard transform, or SRHT (see, e.g., Lemma 6 of [6]) and set $t = O(\varepsilon^{-2}(\log d)(\sqrt{d} + \sqrt{\log n})^2)$, to achieve $O(nd \log t)$ time. These were the fastest known subspace embeddings achieving any value of t not depending polynomially on n . Our main result improves this to achieve $t = \text{poly}(d/\varepsilon)$ for matrices S for which SA can be computed in $\text{nnz}(A)$ time! Given our new subspace embedding, we plug it into known methods of solving the above linear algebra problems given a subspace embedding as a black box.

In fact, our subspace embedding is nothing other than the **CountSketch** matrix in the data stream literature [7], see also [51]. This matrix was also studied by Dasgupta, Kumar, and Sarlós [12]. Formally, S has a single randomly chosen non-zero entry $S_{h(j),j}$ in each column j , for a random mapping $h : [n] \mapsto [t]$. With probability $1/2$, $S_{h(j),j} = 1$, and with probability $1/2$, $S_{h(j),j} = -1$.

While such matrices S have been studied before, the surprising fact is that they actually provide subspace embeddings. Indeed, the usual way of proving that a random $S \sim \pi$ is a subspace embedding is to show that for any fixed vector $y \in \mathbb{R}^d$, $\Pr[\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2] \geq 1 - \exp(-d)$. One then puts a net (see, e.g., [4]) on the unit vectors in the column space $C(A)$, and argues by a union bound that $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all net points y . This then implies, for a net that is sufficiently fine, and using the linearity of the mapping, that $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all vectors $y \in C(A)$.

We stress that our choice of matrices S does not preserve the norms of an arbitrary set of $\exp(d)$ vectors with high probability, and so the above approach cannot work for our choice of matrices S . We instead critically use that these $\exp(d)$ vectors all come from a d -dimensional subspace (namely, $C(A)$), and therefore have a very special structure. The structural fact we use is that there is a fixed set H of size d/α which depends only on the subspace, such that for any unit vector $y \in C(A)$, H contains the indices of all coordinates of y larger than $\sqrt{\alpha}$ in magnitude. The key property here is that the set H is independent of y , or in other words, only a small set of coordinates could ever be large as we range over all unit vectors in the subspace. The set H selects exactly the set of large leverage scores of the columns space $C(A)$!

Given this observation, by setting $t \geq K|H|^2$ for a large enough constant K , we have that with probability $1 - 1/K$, there are no two distinct $j \neq j'$ with $j, j' \in H$ for which $h(j) = h(j')$. That is, we avoid the birthday paradox, and the coordinates in H are “perfectly hashed” with large probability. Call this event \mathcal{E} , which we condition on.

Given a unit vector y in the subspace, we can write it as $y^H + y^L$, where y^H consists of y with the coordinates in $[n] \setminus H$ replaced with 0, while y^L consists of y with the coordinates in H replaced with 0. We seek to bound

$$\|Sy\|_2^2 = \|Sy^H\|_2^2 + \|Sy^L\|_2^2 + 2\langle Sy^H, Sy^L \rangle.$$

Since \mathcal{E} occurs, we have the isometry $\|Sy^H\|_2^2 = \|y^H\|_2^2$. Now, $\|y^L\|_\infty^2 < \alpha$, and so we can apply Theorem 2 of [12] which shows that for mappings of our form, if the input vector has small infinity norm, then S preserves the norm of the vector up to an additive $O(\varepsilon)$ factor with high probability. Here, it suffices to set $\alpha = 1/\text{poly}(d/\varepsilon)$.

Finally, we can bound $\langle Sy^H, Sy^L \rangle$ as follows. Define $G \subseteq [n] \setminus H$ to be the set of coordinates j for which $h(j) = h(j')$ for a coordinate $j' \in H$, that is, those coordinates in $[n] \setminus H$ which “collide” with an element of H . Then, $\langle Sy^H, Sy^L \rangle = \langle Sy^H, Sy^{L'} \rangle$, where $y^{L'}$ is a vector which agrees with y^L on coordinates $j \in G$, and is 0 on the remaining coordinates. By Cauchy-Schwarz, this is at most $\|Sy^H\|_2 \cdot \|Sy^{L'}\|_2$. We have already argued that $\|Sy^H\|_2 = \|y^H\|_2 \leq 1$ for unit vectors y . Moreover, we can again apply Theorem 2 of [12] to bound $\|Sy^{L'}\|_2$, since, conditioned on the coordinates of $y^{L'}$ hashing to the set of items that the coordinates of y^H hash to, they are otherwise random, and so we again have a mapping of our form (with a smaller t and applied to a smaller n) applied to a vector with small infinity-norm. Therefore, $\|Sy^{L'}\|_2 \leq O(\varepsilon) + \|y^{L'}\|_2$ with high probability. Finally, by Bernstein bounds, since the coordinates of y^L are small and t is sufficiently large, $\|y^{L'}\|_2 \leq \varepsilon$ with high probability. Hence, conditioned on event \mathcal{E} , $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ with probability $1 - \exp(-d)$, and we can complete the argument by union-bounding over a sufficiently fine net.

We note that an inspiration for this work comes from work on estimating norms in a data stream with

efficient update time by designing separate data structures for the heavy and the light components of a vector [43, 33]. A key concept here is to characterize the heaviness of coordinates in a vector space in terms of its leverage scores.

Optimizing the additive term: The above approach already illustrates the main idea behind our subspace embedding, providing the first known subspace embedding that can be implemented in $\text{nnz}(A)$ time. This is sufficient to achieve our numerical linear algebra results in time $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ for regression and $O(\text{nnz}(A)) + n \cdot \text{poly}(k \log(n)/\varepsilon)$ for low rank approximation. However, for some applications d, k , or $1/\varepsilon$ may also be large, and so it is important to achieve a small degree in the additive $\text{poly}(d/\varepsilon)$ and $n \cdot \text{poly}(k \log(n)/\varepsilon)$ factors. The number of rows of the matrix S is $t = \text{poly}(d/\varepsilon)$, and the simplest analysis described above would give roughly $t = (d/\varepsilon)^8$. We now show how to optimize this.

The first idea for bringing this down is that the analysis of [12] can itself be tightened by using that we are applying it on vectors coming from a subspace instead of on a set of arbitrary vectors. This involves observing that in the analysis of [12], if on input vector y and for every $i \in [t]$, $\sum_{j|h(j)=i} y_j^2$ is small then the remainder of the analysis of [12] does not require that $\|y\|_\infty$ be small. Since our vectors come from a subspace, it suffices to show that for every $i \in [t]$, $\sum_{j|h(j)=i} \|U_j\|_2^2$ is small, where $\|U_j\|_2^2$ is the j -th leverage score of A . Therefore we do not need to perform this analysis for each y , but can condition on a single event, and this effectively allows us to increase α in the outline above, thereby reducing the size of H , and also the size of t since we have $t = \Omega(|H|^2)$. In fact, we instead follow a simpler and slightly tighter analysis of [32] based on the Hanson-Wright inequality.

Another idea is that the estimation of $\|y^H\|_2$, the contribution from the “heavy coordinates”, is inefficient since it requires a perfect hashing of the coordinates, which can be optimized to reduce the additive term to $d^2 \varepsilon^{-2} \text{polylog}(d/\varepsilon)$. In the worst case, there are d leverage scores of value about 1, $2d$ of value about $1/2$, $4d$ of value about $1/4$, etc. While the top d leverage scores need to be perfectly hashed (e.g., if A contains the $d \times d$ identity matrix as a submatrix), it is not necessary that the leverage scores of smaller value, yet still larger than $1/d$, be perfectly hashed. Allowing a small number of collisions is okay provided all vectors in the subspace have small norm on these collisions, which just corresponds to the spectral norm of a submatrix of A . This gives an additive term of $d^2 \varepsilon^{-2} \text{polylog}(d/\varepsilon)$ instead of $O(d^4 \varepsilon^{-4})$. This refinement is discussed in Section §4.

There is yet another way to optimize the additive term to roughly $d^2(\log n)/\varepsilon^4$, which is useful in its own right since the error probability of the mapping can now be made very low, namely, $1/\text{poly}(n)$. This low error probability bound is needed for our application to ℓ_p -regression, see Section 9. By standard balls-and-bins analyses, if we have $O(d^2/\log n)$ bins and d^2 balls, then with high probability each bin will contain $O(\log n)$ balls. We thus make t roughly $O(d^2/\log n)$ and think of having $O(d^2/\log n)$ bins. In each bin i , $O(\log n)$ heavy coordinates j will satisfy $h(j) = i$. Then, we apply a separate JL transform on the coordinates that hash to each bin i . This JL transform maps a vector $z \in \mathbb{R}^n$ to an $O((\log n)/\varepsilon^2)$ -dimensional vector z' for which $\|z'\|_2 = (1 \pm \varepsilon)\|z\|_2$ with probability at least $1 - 1/\text{poly}(n)$. Since there are only $O(\log n)$ heavy coordinates mapping to a given bin, we can put a net on all vectors on such coordinates of size only $\text{poly}(n)$. We can do this for each of the $O(d^2/\log n)$ bins and take a union bound. It follows that the 2-norm of the vector of coordinates that hash to each bin is preserved, and so the entire vector y^H of heavy coordinates has its 2-norm preserved. By a result of [32], the JL transform can be implemented in $O((\log n)/\varepsilon)$ time, giving total time $O(\text{nnz}(A)(\log n)/\varepsilon)$, and this reduces t to roughly $O(d^2 \log n)/\varepsilon^4$.

We also note that for applications such as least squares regression, it suffices to set ε to be a constant in the subspace embedding, since we can use an approach in [23, 11] which, given constant-factor approximations to all of the leverage scores, can then achieve a $(1 + \varepsilon)$ -approximation to least squares regression by slightly over-sampling rows of the adjointed matrix $A \circ b$ proportional to its leverage scores, and solving the induced subproblem. This results in a better dependence on ε .

We can also compose our subspace embedding with a fast JL transform to further reduce t to the optimal value of about d/ε^2 . Since $S \cdot A$ already has small dimensions, applying a fast JL transform is now efficient.

Finally, we can use a recent result of [8] to replace most dependencies on d in our running times for regression with a dependence on the rank r of A , which may be smaller.

Note that when a matrix A is input that has leverage scores that are roughly equal to each other, then the set H of heavy coordinates is empty. Such a leverage score condition is assumed, for example, in the analysis of matrix completion algorithms. For such matrices, the sketching dimension can be made $d^2\varepsilon^{-2}\log(d/\varepsilon)$, slightly improving our $d^2\varepsilon^{-2}\text{polylog}(d/\varepsilon)$ dimension above.

1.3 Recent Related Work

In the first version of our technical report on these ideas (July, 2012), the additive $\text{poly}(k, d, 1/\varepsilon)$ terms were not optimized, while in the second version, the additive terms were more refined, and results on ℓ_p -regression for general p were given, but the analysis of sparse embeddings in §4 was absent. In the third version, we refined the dependence still further, with the partitioning in §4. Recently, a number of authors have told us of followup work, all building upon our initial technical report.

Miller and Peng showed that ℓ_2 -regression can be done with the additive term sharpened to sub-cubic dependence on d , and with linear dependence on $\text{nnz}(A)$ [41]. More fundamentally, they showed that a subspace embedding can be found in $O(\text{nnz}(A) + d^{\omega+\alpha}\varepsilon^{-2})$ time, to dimension

$$O((d^{1+\alpha}\log d + \text{nnz}(A)d^{-3})\varepsilon^{-2});$$

here ω is the exponent for asymptotically fast matrix multiplication, and $\alpha > 0$ is an arbitrary constant. (Some constant factors here are increasing in α .)

Nelson and Nguyen obtained similar results for regression, and showed that sparse embeddings can embed into dimension $O(d^2/\varepsilon^2)$ in $O(\text{nnz}(A))$ time; this considerably improved on our dimension bound for that running time, at that point (our second version), although our current bound is within $\text{polylog}(d/\varepsilon)$ of their result. They also showed a dimension bound of $O(d^{1+\alpha})$ for $\alpha > 0$, with work $O(f(\alpha)\text{nnz}(A)\varepsilon^{-1})$ for a particular function of α . Their analysis techniques are quite different from ours [42].

Both of these papers use fast matrix multiplication to achieve sub-cubic dependence on d in applications, where our cubic term involves a JL transform, which may have favorable properties in practice. Regarding subspace embeddings to dimensions near-linear in d , note that by computing leverage scores and then sampling based on those scores, we can obtain subspace embeddings to $O(d\varepsilon^{-2}\log d)$ dimensions in $O(\text{nnz}(A)\log n) + \tilde{O}(r^3)$ time; this may be incomparable to the results just mentioned, for which the running times increase as $\alpha \rightarrow 0$, possibly significantly.

Paul, Boutsidis, Magdon-Ismael, and Drineas [46] implemented our subspace embeddings and found that in the TechTC-300 matrices, a collection of 300 sparse matrices of document-term data, with an average of 150 to 200 rows and 15,000 columns, our subspace embeddings as used for the projection step in their SVM classifier are about 20 times faster than the Fast JL Transform, while maintaining the same classification accuracy. Despite this large improvement in the time for projecting the data, further research is needed for SVM classification, as the JL Transform empirically possesses additional properties important for SVM which make it faster to classify the projected data, even though the time to project the data using our method is faster.

Finally, Meng and Mahoney improved on the first version of our additive terms for subspace embeddings, and showed that these ideas can also be applied to ℓ_p -regression, for $1 \leq p < 2$ [39]; our work on this in §9 achieves $1 \leq p < \infty$ and was done independently. We note that our algorithms for ℓ_p -regression require constructions of embeddings that are successful with high probability, as we obtain for generalized embeddings, and so some of the constructions in [41, 42] (as well as our non-generalized embeddings) will not yield such ℓ_p results.

1.4 Outline

We introduce basic notation and definitions in §2, and then the basic analysis in §3. A more refined analysis is given in §4, and then generalized embeddings, with high probability guarantees, in §5. In these sections, we generally follow the framework discussed above, splitting coordinates of column-space vectors into sets of “large” and “small” ones, analyzing each such set separately, and then bringing these analyses together.

Shifting to applications, we discuss leverage score approximation in §6, and regression in §7, including the use of leverage scores and the algorithmic machinery used to estimate them, and considering affine embeddings in §7.5, constrained regression in §7.6, and iterative methods in §7.7. Our low-rank approximation algorithms are given in §8, where we use constructions and analysis based on leverage scores and regression. We next apply generalized sparse embeddings to ℓ_p -regression, in §9. Finally, in §10, we give some preliminary experimental results.

2 Sparse Embedding Matrices

We let $\|A\|_F$ or $\|A\|$ denote the Frobenius norm of matrix A , and $\|A\|_2$ denote the spectral norm of A .

Let $A \in \mathbb{R}^{n \times d}$. We assume $n > d$. Let $\text{nnz}(A)$ denote the number of non-zero entries of A . We can assume $\text{nnz}(A) \geq n$ and that there are no all-zero rows or columns in A .

For a parameter t , we define a random linear map $\Phi D : \mathbb{R}^n \rightarrow \mathbb{R}^t$ as follows:

- $h : [n] \mapsto [t]$ is a random map so that for each $i \in [n]$, $h(i) = t'$ for $t' \in [t]$ with probability $1/t$.
- $\Phi \in \{0, 1\}^{t \times n}$ is a $t \times n$ binary matrix with $\Phi_{h(i), i} = 1$, and all remaining entries 0.
- D is an $n \times n$ random diagonal matrix, with each diagonal entry independently chosen to be $+1$ or -1 with equal probability.

We will refer to a matrix of the form ΦD as a *sparse embedding matrix*.

3 Analysis

Let $U \in \mathbb{R}^{n \times r}$ have columns that form an orthonormal basis for the column space $C(A)$. Let $U_{1,*}, \dots, U_{n,*}$ be the rows of U , and let $u_i \equiv \|U_{i,*}\|^2$.

It will be convenient to regard the rows of A and U to be re-arranged so that the u_i are in non-increasing order, so u_1 is largest; of course this order is unknown and un-used by our algorithms.

For $u \in \mathbb{R}^n$ and $1 \leq a \leq b \leq n$, let $u_{a:b}$ denote the vector with i 'th coordinate equal to u_i when $i \in [a, b]$, and zero otherwise.

Let $T > 0$ be a parameter. Throughout, we let $s \equiv \min\{i | u_i \leq T\}$, and $s' \equiv \max\{i | \sum_{s \leq j \leq i} u_j \leq 1\}$.

We will use the notation $\llbracket P \rrbracket$, a function on event P , that returns 1 when P holds, and 0 otherwise.

The following variation of Bernstein's inequality¹ will be helpful.

Lemma 1 For $L, T \geq 0$ and independent random variables $X_i \in [0, T]$ with $V \equiv \sum_i \text{Var}[X_i]$, if $V \leq LT^2/6$, then

$$\Pr \left[\sum_i X_i \geq \sum_i \mathbf{E}[X_i] + LT \right] \leq \exp(-L).$$

Proof: Here Bernstein's inequality says that for $Y_i \equiv X_i - \mathbf{E}[X_i]$, so that $\mathbf{E}[Y_i^2] = \text{Var}[X_i] = V$ and $|Y_i| \leq T$,

$$\log \Pr \left[\sum_i Y_i \geq z \right] \leq \frac{-z^2/2}{V + zT/3}.$$

By the quadratic formula, the latter is no more than $-L$ when

$$z \geq \frac{LT}{3} (1 + \sqrt{1 + 18V/LT^2}),$$

which holds for $z \geq LT$ and $V \leq LT^2/6$. ■

¹See Wikipedia entry on Bernstein's inequalities (probability theory).

3.1 Handling vectors with small entries

We begin the analysis by considering $y_{s:n}$ for fixed unit vectors $y \in C(A)$. Since $\|y\| = 1$, there must be a unit vector x so that $y = Ux$, and so by Cauchy-Schwartz, $\|y_i\|^2 \leq \|U_{i,*}\|^2 \|x\|^2 = u_i$. This implies that $\|y_{s:n}\|_\infty^2 \leq u_s$. We extend this to all unit vectors in subsequent sections.

The following is similar to Lemma 6 of [12], and is a standard balls-and-bins analysis.

Lemma 2 For $\delta_h, T, t > 0$, and $s \equiv \min\{i \mid u_i \leq T\}$, let \mathcal{E}_h be the event that

$$W \geq \max_{j \in [t]} \sum_{\substack{i \in h^{-1}(j) \\ i \geq s}} u_i,$$

where $W \equiv T \log(t/\delta_h) + r/t$. If

$$t \geq \frac{6\|u_{s:n}\|^2}{T^2 \log(t/\delta_h)},$$

then $\Pr[\mathcal{E}_h] \geq 1 - \delta_h$.

Proof: We will apply Lemma 1 to prove that the bound holds for fixed $j \in [t]$ with failure probability δ_h/t , and then apply a union bound.

Let X_i denote the random variable $u_i \mathbb{I}[h(i) = j, i \geq s]$. We have $0 \leq X_i \leq T$, $\mathbf{E}[X] = \sum_{i \geq s} u_i/t \leq r/t$, and $V = \sum_{i \geq s} \mathbf{E}[X_i^2] = \sum_{i \geq s} u_i^2/t = \|u_{s:n}\|^2/t$. Applying Lemma 1 with $L = \log(t/\delta_h)$ gives

$$\Pr\left[\sum_i X_i \geq T \log(t/\delta_h) + r/t\right] \leq \exp(-\log(t/\delta_h)) = \delta_h/t,$$

when $\|u_{s:n}\|^2/t \leq LT^2/6$, or $t \geq 6\|u_{s:n}\|^2/LT^2$. ■

Lemma 3 For W as in Lemma 2, suppose the event \mathcal{E}_h holds. Then for unit vector $y \in C(A)$, and any $2 \leq \ell \leq 1/W$, with failure probability $\delta_L = e^{-\ell}$, $|\|\Phi D y_{s:n}\|_2^2 - \|y_{s:n}\|^2| \leq K_L \sqrt{W \log(1/\delta_L)}$, where K_L is an absolute constant.

Proof: We will use the following theorem, due to Hanson and Wright.

Theorem 4 [30] Let $z \in \mathbb{R}^n$ be a vect or of i.i.d. ± 1 random values. For any symmetric $B \in \mathbb{R}^{n \times n}$ and $2 \leq \ell$,

$$\mathbf{E} [|z^\top B z - \text{tr}(B)|^\ell] \leq (CQ)^\ell,$$

where

$$Q \equiv \max\{\sqrt{\ell}\|B\|_F, \ell \cdot \|B\|_2\},$$

and $C > 0$ is a universal constant.

We will use Theorem 4 to prove a bound on the ℓ 'th moment of $\|\Phi D y\|_2^2$ for large ℓ . Note that $\|\Phi D y\|_2^2$ can be written as $z^\top B z$, where z has entries from the diagonal of D , and $B \in \mathbb{R}^{n \times n}$ has $B_{ii'} \equiv y_i y_{i'} \mathbb{I}[h(i) = h(i')]$. Here $\text{tr}(B) = \|y_{s:n}\|^2$.

Our analysis uses some ideas from the proofs for Lemmas 7 and 8 of [32].

Since by assumption event \mathcal{E}_h of Lemma 2 occurs, and for unit $y \in C(A)$, $y_{i'}^2 \leq u_{i'}$ for all i' , we have for $j \in [t]$ that $\sum_{i' \in h^{-1}(j), i' \geq s} y_{i'}^2 \leq W$. Hence

$$\begin{aligned} \|B\|_F^2 &= \sum_{i, i' \geq s} (y_i y_{i'})^2 \mathbb{I}[h(i') = h(i)] \\ &= \sum_{i \geq s} y_i^2 \sum_{\substack{i' \in h^{-1}(h(i)) \\ i' \geq s}} y_{i'}^2 \\ &\leq \sum_{i \in [n]} y_i^2 W \\ &\leq W. \end{aligned} \tag{1}$$

For $\|B\|_2$, observe that for given $j \in [t]$, $z(j) \in \mathbb{R}^n$ with $z(j)_i = y_i \mathbb{I}[h(i) = j, i \geq s]$ is an eigenvector of B with eigenvalue $\|z(j)\|^2$, and the set of such eigenvectors spans the column space of B . It follows that

$$\|B\|_2 = \max_j \|z(j)\|^2 = \sum_{\substack{i' \in h^{-1}(j) \\ i' \geq s}} y_{i'}^2 \leq W.$$

Putting this and (1) into the Q of Theorem 4, we have,

$$Q \leq \max\{\sqrt{\ell}\|B\|_F, \ell \cdot \|B\|_2\} \leq \max\{\sqrt{\ell}\sqrt{\ell W}, \ell W\} = \sqrt{\ell W},$$

where we used $\ell W \leq 1$. By a Markov bound applied to $|z^\top Bz - \text{tr}(B)|^\ell$ with $\ell = \log(1/\delta_L)$,

$$\begin{aligned} \Pr[|\|\Phi D y_{s:n}\|_2^2 - \|y_{s:n}\|^2| \geq eC\sqrt{\ell W}] &\leq e^{-\ell} \\ &= \delta_L. \end{aligned}$$

■

3.2 Handling vectors with large entries

A small number of entries can be handled directly.

Lemma 5 *For given s , let \mathcal{E}_B denote the event that $h(i) \neq h(i')$ for all $i, i' < s$. Then $\delta_B \equiv 1 - \Pr[\mathcal{E}_B] \leq s^2/t$. Given event \mathcal{E}_B , we have that for any y ,*

$$\|y_{1:(s-1)}\|_2^2 = \|\Phi D y_{1:(s-1)}\|_2^2.$$

Proof: Since $\Pr[h(i) = h(i')] = 1/t$, the probability that some such $i \neq i'$ has $h(i) = h(i')$ is at most s^2/t . The last claim follows by a union bound. ■

3.3 Handling all vectors

We have seen that ΦD preserves the norms for vectors with small entries (Lemma 3) and large entries (Lemma 5). Before proving a general bound, we need to prove a bound on the “cross terms”.

Lemma 6 *For W as in Lemma 2, suppose the event \mathcal{E}_h and \mathcal{E}_B hold. Then for unit vector $y \in C(A)$, with failure probability at most δ_C ,*

$$|y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n}| \leq K_C \sqrt{W \log(1/\delta_C)},$$

for an absolute constant K_C .

Proof: With the event \mathcal{E}_B , for each $i \geq s$ there is at most one $i' < s$ with $h(i) = h(i')$; let $z_i \equiv y_{i'} D_{i'i'}$, and $z_i \equiv 0$ otherwise. We have for integer $p \geq 1$ using Khintchine's inequality

$$\begin{aligned} \mathbf{E} \left[\left(y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n} \right)^{2p} \right]^{1/p} &= \mathbf{E} \left[\left(\sum_{i \geq s} y_i D_{ii} z_i \right)^{2p} \right]^{1/p} \\ &\leq C_p \sum_{i \geq s} y_i^2 z_i^2 \\ &= C_p \sum_{i' < s} y_{i'}^2 \sum_{\substack{i \in h^{-1}(i') \\ i \geq s}} y_i^2 \\ &\leq C_p W, \end{aligned}$$

where $C_p \leq \Gamma(p+1/2)^{1/p} = O(p)$, and the last inequality uses the assumption that \mathcal{E}_h holds, and $\sum_{i' < s} y_{i'}^2 \leq 1$. Putting $p = \log(1/\delta_C)$ and applying the Markov inequality, we have

$$\Pr[(y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n})^2 \geq e C_p W] \geq 1 - \exp(-p) = 1 - \delta_C.$$

Therefore, with failure probability at most δ_C , we have

$$|y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n}| \leq K_C \sqrt{W \log(1/\delta_C)},$$

for an absolute constant K_C . ■

Lemma 7 Suppose the events \mathcal{E}_h and \mathcal{E}_B hold, and W is as in Lemma 2. Then for $\delta_y > 0$ there is an absolute constant K_y such that, if $W \leq K_y \epsilon^2 / \log(1/\delta_y)$, then for unit vector $y \in C(A)$, with failure probability δ_y , $\|\Phi D y\|_2 = (1 \pm \epsilon)\|y\|_2$, when $\delta_y \leq 1/2$.

Proof: Assuming \mathcal{E}_h and \mathcal{E}_B , we apply Lemmas 5, 3, and 6, and have with failure probability at most $\delta_L + \delta_C$,

$$\begin{aligned} &| \|\Phi D y\|_2^2 - \|y\|_2^2 | \\ &= | \|\Phi D y_{1:(s-1)}\|_2^2 - \|y_{1:(s-1)}\|_2^2 \\ &\quad + \|\Phi D y_{s:n}\|_2^2 - \|y_{s:n}\|_2^2 + 2 y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n} | \\ &\leq | \|\Phi D y_{s:n}\|_2^2 - \|y_{s:n}\|_2^2 | + 0 + |2 y_{1:(s-1)}^\top D \Phi^\top \Phi D y_{s:n}| \\ &\leq K_L \sqrt{W \log(1/\delta_L)} + 2 K_C \sqrt{W \log(1/\delta_C)} \\ &\leq 3\epsilon \sqrt{K_y} (K_L + K_C) \end{aligned}$$

for the given W , putting $\delta_L = \delta_C = \delta_y/2$ and assuming $\delta_y \leq 1/2$. Thus $K_y \leq 1/9(K_L + K_C)^2$ suffices. ■

Lemma 8 Suppose $\delta_{sub} > 0$, L is an r -dimensional subspace of \mathbb{R}^n , and $B : \mathbb{R}^n \rightarrow \mathbb{R}^k$ is a linear map. If for any fixed $x \in L$, $\|Bx\|_2^2 = (1 \pm \epsilon/6)\|x\|_2^2$ with probability at least $1 - \delta_{sub}$, then there is a constant $K_{sub} > 0$ for which with probability at least $1 - \delta_{sub} K_{sub}^r$, for all $x \in L$, $\|Bx\|_2^2 = (1 \pm \epsilon)\|x\|_2^2$.

Proof: We will need the following standard lemmas for making a net argument. Let S^{r-1} be the unit sphere in \mathbb{R}^r and let E be the set of points in S^{r-1} defined by

$$E = \left\{ w : w \in \frac{\gamma}{\sqrt{r}} \mathbb{Z}^r, \|w\|_2 \leq 1 \right\},$$

where \mathbb{Z}^r is the r -dimensional integer lattice and γ is a parameter.

Fact 9 (Lemma 4 of [4]) $|E| \leq e^{cr}$ for $c = (\frac{1}{\gamma} + 2)$.

Fact 10 (Lemma 4 of [4]) For any $r \times r$ matrix J , if for every $u, v \in E$ we have $|u^\top Jv| \leq \varepsilon$, then for every unit vector w , we have $|w^\top Jw| \leq \frac{\varepsilon}{(1-\gamma)^2}$.

Let $U \in \mathbb{R}^{n \times r}$ be such that the columns are orthonormal and the column space equals L . Let I_r be the $r \times r$ identity matrix. Define $J = U^T B^T B U - I_r$. Consider the set E in Fact 9 and Fact 10. Then, for any $x, y \in E$, we have by the statement of the lemma that with probability at least $1 - 3\delta_{sub}$, $\|B U x\|_2^2 = (1 \pm \varepsilon/6)\|U x\|_2^2$, $\|B U y\|_2^2 = (1 \pm \varepsilon/6)\|U y\|_2^2$, and $\|B U(x+y)\|_2^2 = (1 \pm \varepsilon/6)\|U(x+y)\|_2^2 = (1 \pm \varepsilon/6)(\|U x\|_2^2 + \|U y\|_2^2 + 2\langle U x, U y \rangle)$. Since $\|U x\|_2 \leq 1$ and $\|U y\|_2 \leq 1$, it follows that $|x J y| \leq \varepsilon/2$. By Fact 9, for $\gamma = 1 - 1/\sqrt{2}$ and sufficiently large K_{sub} , we have by a union bound that with probability at least $1 - \delta_{sub} K_{sub}^r$ that $|x J y| \leq \varepsilon/2$ for every $x, y \in E$. Hence, with this probability, by Fact 10, $|w^\top J w| \leq \varepsilon$ for every unit vector w , which by definition of J means that for all $y \in L$, $\|B y\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$. \blacksquare

The following is our main theorem in this section.

Theorem 11 There is $t = O((r/\epsilon)^4 \log^2(r/\epsilon))$ such that with probability at least $9/10$, ΦD is a subspace embedding matrix for A ; that is, for all $y \in C(A)$, $\|\Phi D y\|_2 = (1 \pm \varepsilon)\|y\|_2$. The embedding ΦD can be applied in $O(\text{nnz}(A))$ time. For $s = \min\{i' \mid u_{i'} \leq T\}$, where T is a parameter in $\Omega(\epsilon^2/r \log(r/\epsilon))$, it suffices if $t \geq \max\{s^2/30, r/T\}$.

Proof: For suitable t , T , and s , with failure probability at most $\delta_h + \delta_B$, events \mathcal{E}_h and \mathcal{E}_B both hold. Conditioned on this, and assuming W is sufficiently small as in Lemma 7, we have with failure probability δ_y for any fixed $y \in C(A)$ that $\|\Phi D y\|_2 = (1 \pm \varepsilon)\|y\|_2$. Hence by Lemma 8, with failure probability $\delta_h + \delta_B + \delta_y K_{sub}^r$, $\|\Phi D y\|_2 = (1 \pm 6\varepsilon)\|y\|_2$ for all $y \in C(A)$. We need $\delta_h + \delta_B + \delta_y K_{sub}^r \leq 1/10$, and the parameter conditions of Lemmas 2, Lemma 3, and Lemma 7 holding. Listing these conditions:

1. $\delta_h + \delta_B + \delta_y K_{sub}^r \leq 1/10$, where δ_B can be set to be s^2/t ;
2. $u_s \leq T$;
3. $t \geq 6\|u_{s:n}\|^2 / \log(t/\delta_h) T^2$;
4. $\ln(2/\delta_y) \cdot W \leq 1$ (corresponding to the condition $\ell \leq 1/W$ of Lemma 3 since we set $\delta_y/2 = \delta_L = e^{-\ell}$)
5. $W = T \log(t/\delta_h) + r/t \leq K_y \epsilon^2 / \log(1/\delta_y)$.

We put $\delta_y = K_{sub}^{-r}/30$, $\delta_h = 1/30$, and require $t \geq s^2/30$. For the last condition it suffices that $T = O(\epsilon^2/r \log(t))$, and $t = \Omega(r^2/\epsilon^2)$. The last condition implies the fourth condition for small enough constant ε . Also, since $\|u_{s:n}\|^2 = \sum_{i \geq s} u_i^2 \leq \sum_{i \geq s} u_i T \leq rT$, the bound for T implies that $t = O((r/\epsilon)^2 \log(t))$ suffices for Condition 3. Thus when the leverage scores are such that s is small, t can be $O((r/\epsilon)^2 \log(r/\epsilon))$. Since $\sum_i u_i = r$, $s \leq r/T$ suffices, and so $t = O((r/T)^2) = O((r/\epsilon)^4 \log^2(r/\epsilon))$ suffices for the conditions of the theorem. \blacksquare

4 Partitioning Leverage Scores

We can further optimize our low order additive $\text{poly}(r)$ term by refining the analysis for large leverage scores (those larger than T). We partition the scores into groups that are equal up to a constant factor, and analyze the error resulting from the relatively small number of collisions that may occur, using also the leverage scores to bound the error. In what follows we have not optimized the $\text{poly}(\log(r/\varepsilon))$ factors.

Let $q \equiv \log_2 1/T = O(\log(r/\varepsilon))$. We partition the leverage scores u_i with $u_i \geq T$ into groups G_j , $j \in [q]$, where

$$G_j = \{i \mid 1/2^j < u_i \leq 1/2^{j-1}\}.$$

Let $\beta_j \equiv 2^{-j}$, and $n_j \equiv |G_j|$. Since $\sum_{i=1}^n u_i = r$, we have for all j that $n_j \leq r/\beta_j$.

We may also use G_j to refer to the collection of rows of U with leverage scores in G_j .

For given hash function h and corresponding Φ , let $G'_j \subset G_j$ denote the collision indices of G_j , those $i \in G_j$ such that $h(i) = h(i')$ for some $i' \in G_j$. Let $k_j \equiv |G'_j|$.

First, we bound the spectral norm of a submatrix of the orthogonal basis U of $C(A)$, where the submatrix comprises rows of G'_j .

4.1 The Spectral Norm

We have a matrix $B \in \mathbb{R}^{n_j \times r}$, with $\|B\|_2 \leq 1$, and each row of B has squared Euclidean norm at least β_j and at most $2\beta_j$, for some $j \in [q]$.

We want to bound the spectral norm of the matrix \hat{B} whose rows comprise those rows of U in the collision set G'_j . We let $t = \Theta(r^2 q^6 / \epsilon^2)$ be the number of hash buckets. The expected number of collisions in the t buckets is $\mathbf{E}[|G'_j|] = \frac{\binom{n_j}{2}}{t} \leq \frac{n_j^2}{2t}$. Let \mathcal{D}_j be the event that the number $k_j \equiv |G'_j|$ of such collisions in the t buckets is at most $n_j^2 q^2 / t$. Let $\mathcal{D} = \cap_{j=1}^q \mathcal{D}_j$. By a Markov and a union bound, $\Pr[\mathcal{D}] \geq 1 - 1/(2q)$. We will assume that \mathcal{D} occurs.

While each row in B has some independent probability of participating in a collision, we first analyze a sampling scheme with replacement.

We generate independent random matrices \hat{H}_m for $m \in [\ell_j]$, for a parameter $\ell_j > k_j$, by picking $i \in [n_j]$ uniformly at random, and letting $\hat{H}_m \equiv B_{i:}^\top B_{i:}$. Note that $\mathbf{E}[\hat{H}_m] = \frac{1}{n_j} B^\top B$.

Our analysis will use a special case of the version of matrix Bernstein inequalities described by Recht.

Fact 12 (paraphrase of Theorem 3.2 [47]) *Let ℓ be an integer parameter. For $m \in [\ell]$, let $H_m \in \mathbb{R}^{r \times r}$ be independent symmetric zero-mean random matrices. Suppose $\rho_m^2 \equiv \|\mathbf{E}[H_m H_m]\|_2$ and $M \equiv \max_{m \in [\ell]} \|H_m\|_2$. Then for any $\tau > 0$,*

$$\log \Pr \left[\left\| \sum_{m \in [\ell]} H_m \right\|_2 > \tau \right] \leq \log 2r - \frac{\tau^2/2}{\sum_{m \in [\ell]} \rho_m^2 + M\tau/3}.$$

We apply this fact with $\ell = \ell_j = (4e^2)k_j + \Theta(q)$ and $H_m \equiv \hat{H}_m - \mathbf{E}[\hat{H}_m]$, so that

$$\begin{aligned} \rho_m^2 &\equiv \|\mathbf{E}[H_m H_m]\|_2 \\ &\leq \left\| \frac{1}{n_j} \sum_{i \in [n_j]} \|B_{i:}\|^2 B_{i:}^\top B_{i:} - \frac{1}{n_j^2} B^\top B B^\top B \right\|_2 \\ &\leq \frac{2\beta_j}{n_j} + \frac{1}{n_j^2}. \end{aligned}$$

Also $M \equiv \|H_m\|_2 \leq 2\beta_j + \frac{1}{n_j}$.

Applying the above fact with these bounds for ρ_m^2 and M , we have

$$\begin{aligned} \log \Pr \left[\left\| \sum_{m \in [\ell_j]} H_m \right\|_2 > \tau \right] &\leq \log 2r - \frac{\tau^2/2}{\sum_{m \in [\ell_j]} \rho_m^2 + M\tau/3} \\ &\leq \log 2r - \frac{\tau^2/2}{(2\beta_j + 1/n_j) \left(\frac{\ell_j}{n_j} + \frac{\tau}{3} \right)}. \end{aligned}$$

We will assume that $n_j \geq \sqrt{t}/q^2$, as discussed in lemma 13 below (otherwise we have perfect hashing). With this assumption, setting $\tau = \Theta(q(\beta_j + 1/n_j + \sqrt{r/t}))$ gives a probability bound of $1/r$. (Here we use that

$\beta_j + 1/n_j \leq 2$, $\ell_j/n_j = O(qn_j/t)$, and $\frac{n_j}{t}(\beta_j + 1/n_j) \leq (r+1)/t$.) We therefore have that with probability at least $1 - 1/r$,

$$\begin{aligned} \left\| \sum_{m \in [\ell_j]} \hat{H}_m \right\|_2 &= O(q(\beta_j + 1/n_j + \sqrt{r/t}) + \frac{\ell_j}{n_j} \|B^\top B\|_2) \\ &= O(q(\beta_j + 1/n_j + \sqrt{r/t} + \frac{n_j}{t})), \end{aligned}$$

where we use that $\|B\|_2 \leq 1$, and use again $\ell_j/n_j = O(qn_j/t)$.

We can now prove the following lemma.

Lemma 13 *With probability $1 - o(1)$, for all leverage score groups G_j , and for U an orthonormal basis of $C(A)$, the submatrix \hat{B}_j of U consisting of rows in G'_j , that is, those in G_j that collide in a hash bucket with another row in G_j under Φ , has squared spectral norm $O(q(\beta_j + 1/n_j + \sqrt{r/t} + n_j/t))$.*

Proof: Fix a $j \in [q]$. If $n_j \equiv |G_j| \leq \sqrt{t}/q^2$, then with probability $1 - o(1/q)$, the items in G_j are perfectly hashed into the t bins. So with probability $1 - o(1)$, for all $j \in [q]$, if $n_j \leq \sqrt{t}/q^2$, then there are no collisions. Condition on this event.

Now consider a $j \in [q]$ for which $n_j \geq \sqrt{t}/q$. Then

$$\ell_j = (4e^2)k_j + \Theta(q) \leq n_j^2/t + O(q) \leq n_j + O(q) \leq 2n_j.$$

When sampling with replacement, the expected number of distinct items is

$$n_j \cdot \binom{\ell_j}{1} \frac{1}{n_j} \left(1 - \frac{1}{n_j}\right)^{\ell_j-1} \geq \ell_j(1 - o(1))/e^2.$$

By a standard application of Azuma's inequality, using that $\ell_j = \Omega(q)$ is sufficiently large, we have that the number of distinct items is at least $\ell_j/(4e^2)$ with probability at least $1 - 1/r$. By a union bound, with probability $1 - o(1)$, for all $j \in [q]$, if $n_j \geq r$, then at least $\ell_j/(4e^2)$ distinct items are sampled when sampling ℓ_j items with replacement from G_j . Since $\ell_j = 4e^2k_j + O(q)$, it follows that at least k_j distinct items are sampled from each G_j .

By the analysis above, for a fixed $j \in [q]$ we have that the submatrix of U consisting of the ℓ_j sampled rows in G_j has squared spectral norm $O(q(\beta_j + 1/n_j + \sqrt{r/t} + n_j/t))$ with probability at least $1 - 1/r$ (notice that $\|\sum_{m \in [\ell_j]} \hat{H}_m\|_2$ is the square of the spectral norm of the submatrix of U consisting of the ℓ_j sampled rows from G_j). Since the probability of this event is at least $1 - 1/r$ for a fixed $j \in [q]$, we can conclude that it holds for all $j \in [q]$ simultaneously with probability $1 - o(1)$. Finally, using that the spectral norm of a submatrix of a matrix is at most that of the matrix, we have that for each j , the squared spectral norm of a submatrix of k_j random distinct rows among the ℓ_j sampled rows of G_j from U is at most $O(q(\beta_j + 1/n_j + \sqrt{r/t} + n_j/t))$. \blacksquare

4.2 Within-Group Errors

Let $L_j \subset \mathbb{R}^n$ denote the set of vectors y so that $y_i = 0$ for i not in the collision set G'_j , and there is some unit $y' \in C(A)$ such that $y_i = y'_i$ for $i \in G'_j$. (Note that the error for such vectors is that same as that for the corresponding set of vectors with zeros outside of G'_j .)

In this subsection, we show that for all $y \in L_j$, the error in estimating $\|y\|^2$ using $y^\top D\Phi^\top \Phi D y$ is at most $O(\varepsilon)$.

For $y \in L_j$, the error in estimating $\|y\|^2$ by using $y^\top D\Phi^\top \Phi D y$ contributed by collisions among coordinates y_i for $i \in G_j$ is

$$\kappa_j \equiv \sum_{t' \in [t]} \sum_{i, i' \in h^{-1}(t') \cap G_j} y_i y_{i'} D_{ii} D_{i'i'}, \quad (2)$$

and we need a bound on this quantity that holds with high probability.

By a standard balls-and-bins analysis, every bucket has $O(\log t) = O(q)$ collisions, with high probability, since $n_j \leq r/T \leq O(r^2/\varepsilon^2) = O(t)$; we assume this event.

The squared Euclidean norm of the vector of all y_i that appear in the summands, that is, with $i \in G'_j$, is at most $\beta_j + 1/n_j + \sqrt{r/t} + n_j/t$ by Lemma 13. Thus the squared Euclidean norm of the vector comprising all summands in (2) is at most

$$\gamma_j \equiv \sum_{t' \in [t]} \sum_{i, i' \in h^{-1}(t') \cap G_j} y_i^2 y_{i'}^2 \quad (3)$$

$$\begin{aligned} &\leq \sum_{t' \in [t]} \sum_{i \in h^{-1}(t') \cap G_j} y_i^2 O(q) 2\beta_j \\ &\leq O(q^2 \beta_j (\beta_j + 1/n_j + \sqrt{r/t} + n_j/t)). \end{aligned} \quad (4)$$

By Khintchine's inequality, for $p \geq 1$,

$$\mathbf{E}[\kappa_j^{2p}]^{1/p} \leq O(p) \gamma_j \leq O(p) (q^2 \beta_j (\beta_j + 1/n_j + \sqrt{r/t} + n_j/t)),$$

and therefore $|\kappa_j|^2$ is less than the last quantity, with failure probability at most 4^{-p} .

Putting $p = k'_j \equiv \min\{r, k_j\}$, with failure probability at most $4^{-k'_j}$, for any fixed vector $y \in L_j$, the squared error in estimating $\|y\|^2$ using the sketch of y is at most $O(k'_j (q^2 \beta_j (\beta_j + 1/n_j + \sqrt{r/t} + n_j/t)))$. Assuming the event \mathcal{D} from the section above, we have $k'_j \leq \min\{r, q \cdot n_j^2 q/t\}$. We have, using $\beta_j n_j \leq r$,

$$\frac{n_j^2 q^2}{t} (q^2 \beta_j (\beta_j + 1/n_j)) \leq \frac{q^4 r (r+1)}{t},$$

and $r \cdot q^2 \beta_j n_j/t \leq q^2 r^2/t$, and finally

$$q^2 \beta_j \sqrt{r/t} \min\{r, q^2 n_j^2/t\} \leq q^3 \sqrt{r/t} \min\{\beta_j r, \frac{qr^2}{\beta_j t}\} \leq q^4 r^2/t,$$

using $\beta_j n_j \leq r$. Putting these bounds on the terms together, the squared error is $O(q^4 r^2/t)$, or ϵ^2/q^2 , for $t = \Omega(q^6 r^2/\epsilon^2)$, so that the error is $O(\epsilon/q)$.

Since the dimension of L_j is bounded by k'_j , it follows from the net argument of Lemma 8 that for all $y \in L_j$, $\|Sy\|^2 = \|y\|^2 \pm O(\epsilon/q)$, and so the total error for unit $y \in C(A)$ is $O(\epsilon)$.

We thus have the following theorem.

Theorem 14 *There is an absolute constant $C' > 0$ for which for any parameters $\delta_1 \in (0, 1)$, $P \geq 1$, and for sparse embedding dimension $t = O(P(r/\varepsilon)^2 \log^6(r/\varepsilon))$, for all unit $y \in C(A)$, $\sum_{j \in [q]} \|Sy^j\| = 1 \pm C'\epsilon/P\delta_1$, with failure probability at most $\delta_1 + O(1/\log r)$, where y^j denotes the member of L_j derived from y .*

4.3 Handling the Cross Terms

To complete the optimization, we must also handle the error due to “cross terms”.

Let $\delta_1 \in (0, 1)$ be an arbitrary parameter. For $j \neq j' \in \{1, \dots, q\}$, let the event $\mathcal{E}_{j,j'}$ be that the number of bins containing both an item in G_j and in $G_{j'}$ is at most $\frac{n_j n_{j'} q^2}{t \delta_1}$. Let $\mathcal{E} = \cap_{j,j'} \mathcal{E}_{j,j'}$, the event that no pair of groups has too many inter-group collisions.

Lemma 15 $\Pr[\mathcal{E}] \geq 1 - \delta_1$.

Proof: Fix a $j \neq j' \in \{1, \dots, q\}$. Then the expected number of bins containing an item in both G_j and in $G_{j'}$ is at most $t \cdot \frac{n_j}{t} \cdot \frac{n_{j'}}{t} = \frac{n_j n_{j'}}{t}$, and so by a Markov bound the number of bins containing an item in both G_j and $G_{j'}$ is at most $\frac{n_j n_{j'} q^2}{t \delta_1}$ with probability at least $1 - \delta_1/q^2$. The lemma follows by a union bound over the $\binom{q}{2}$ choices of j, j' . \blacksquare

In the remainder of the analysis, we set $t = P(r/\varepsilon)^2 q^6$ for a parameter $P \geq 1$.

Let \mathcal{F} be the event that no bin contains more than Cq elements of $\cup_{i=1}^q G_j$, where $C > 0$ is an absolute constant.

Lemma 16 $\Pr[\mathcal{F}] \geq 1 - 1/r$.

Proof: Observe that $|\cup_{i=1}^q G_j| = \sum_{i=1}^q n_j \leq r \sum_{i=1}^q 2^j \leq 2r^2/\varepsilon^2$. By standard balls and bins analysis with the given t , with $P \geq 1$, with probability at least $1 - 1/r$ no bin contains more than Cq elements, for a constant $C > 0$. ■

Lemma 17 *Condition on events \mathcal{E} and \mathcal{F} occurring. Consider any unit vector $y = Ax$ in the column space of A . Consider any $j \neq j' \in [q]$. Define the vector y^j : $y_i^j = y_i$ for $i \in G_j$, and $y_i^j = 0$ otherwise. Then,*

$$|\langle Sy^j, Sy^{j'} \rangle| = O\left(\frac{1}{P\delta_1 q^2}\right).$$

Proof: Since \mathcal{E} occurs, the number of bins containing both an item in G_j and $G_{j'}$ is at most $n_j n_{j'} q^2 / (t\delta_1)$. Call this set of bins S . Moreover, since \mathcal{F} occurs, for each bin $i \in S$, there are at most $C \log r$ elements from G_j in the bin and at most $C \log r$ elements from $G_{j'}$ in the bin. Hence, for any $S = \Phi \cdot D$, we have, using $n_j \beta_j \leq r$ for all j ,

$$|\langle Sy^j, Sy^{j'} \rangle| \leq \frac{n_j n_{j'} q^2}{t\delta_1} \cdot (Cq)^2 \beta_j \beta_{j'} \leq \frac{(Cq)^2 d^2 q^2}{t\delta_1} = \frac{C^2}{P\delta_1 q^2}.$$

■

The following is our main theorem concerning cross-terms in this section.

Theorem 18 *There is an absolute constant $C' > 0$ for which for any parameters $\delta_1 \in (0, 1)$, $P \geq 1$, and for sparse embedding dimension $t = O(P(r/\varepsilon)^2 \log^6 r)$, the event*

$$\forall y = Ax \text{ with } \|y\|_2 = 1, \sum_{j, j' \in [q]} |\langle Sy^j, Sy^{j'} \rangle| \leq \frac{C\varepsilon^2}{P\delta_1}$$

occurs with failure probability at most $\delta_1 + \frac{1}{r}$, where $y^j, y^{j'}$ are as defined in Lemma 17.

Proof: The theorem follows at once by combining Lemma 15, Lemma 16, and Lemma 17. ■

4.4 Putting it together

Putting the bounds for within-group and cross-term errors together, and replacing the use of Lemma 5 in the proof of Theorem 11, we have the following theorem.

Theorem 19 *There is an absolute constant $C' > 0$ for which for any parameters $\delta_1 \in (0, 1)$, $P \geq 1$, and for sparse embedding dimension $t = O(P(r/\varepsilon)^2 \log^6(r/\varepsilon))$, for all unit $y \in C(A)$, $\|Sy\| = 1 \pm C'\varepsilon/P\delta_1$, with failure probability at most $\delta_1 + O(1/\log r)$.*

5 Generalized Sparse Embedding Matrices

5.1 Johnson-Lindenstrauss transforms

We start with a theorem of Kane and Nelson [32], restated here in our notation. We also present a simple corollary that we need concerning very low dimensional subspaces. Let $\varepsilon > 0$, $a = \Theta(\varepsilon^{-1} \log(r/\varepsilon))$, and

$v = \Theta(\varepsilon^{-1})$. Let $B : \mathbb{R}^n \rightarrow \mathbb{R}^{va}$ be defined as follows. We view B as the concatenation (meaning, we stack the rows on top of each other) of matrices $\sqrt{1/a} \cdot \Phi_1 \cdot D_1, \dots, \sqrt{1/a} \cdot \Phi_a \cdot D_a$, each $\Phi_i \cdot D_i$ being a linear map from \mathbb{R}^n to \mathbb{R}^v , which is an independently chosen sparse embedding matrix of Section 3 with associated hash function $h_i : [n] \rightarrow [v]$.

Theorem 20 ([32]) *For any $\delta_{KN}, \varepsilon > 0$, there are $a = \Theta(\varepsilon^{-1} \log(1/\delta_{KN}))$ and $v = \Theta(\varepsilon^{-1})$ for which for any fixed $x \in \mathbb{R}^n$, a randomly chosen B of the form above satisfies $\|Bx\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2$ with probability at least $1 - \delta_{KN}$.*

Corollary 21 *Let $\delta \in (0, 1)$. Suppose L is an $O(\log(r/\varepsilon\delta))$ -dimensional subspace of \mathbb{R}^n . Let $C_{subKN} > 0$ be any constant. Then for any $\varepsilon \in (0, 1)$, there are $a = \Theta(\varepsilon^{-1} \log(r/\varepsilon\delta))$ and $v = \Theta(\varepsilon^{-1})$ such that with failure probability at most $(\varepsilon/r\delta)^{C_{subKN}}$, $\|By\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$ for all $y \in L$.*

Proof: We use Theorem 20 together with Lemma 8; for the latter, we need that for any fixed $y \in L$, $\|By\|_2^2 = (1 \pm \varepsilon/6)\|y\|_2^2$ with probability at least $1 - \delta_{sub}$. By Theorem 20, we have this for $\delta_{sub} = (\delta\varepsilon/r)^{C_{KN}}$ for an arbitrarily large constant $C_{KN} > 0$. Hence, by Lemma 8, there is a constant $K_{sub} > 0$ so that with probability at least $1 - (K_{sub})^{O(\log(r/\varepsilon\delta))}(\delta\varepsilon/r)^{C_{KN}} = 1 - (\delta\varepsilon/r)^{C_{subKN}}$, for all $y \in L$, $\|By\|_2^2 = (1 \pm \varepsilon)\|y\|_2^2$. Here we use that $C_{KN} > 0$ can be made arbitrarily large, independent of K_{sub} . ■

5.2 The construction

We now define a *generalized sparse embedding matrix* S . Let $A \in \mathbb{R}^{n \times d}$ with rank r .

Let $a = \Theta(\varepsilon^{-1} \log(r/\varepsilon\delta))$ and $v = \Theta(\varepsilon^{-1})$, be such that Theorem 20 and Corollary 21 apply with parameters a and v , for a sufficiently large constant $C_{subKN} > 0$. Further, let

$$q \equiv C_t r \varepsilon^{-2} (r + \log(1/\delta\varepsilon)),$$

where $C_t > 0$ is a sufficiently large absolute constant, and let $t \equiv avq$.

Let $h : [n] \rightarrow [q]$ be a random hash function. For $i = 1, 2, \dots, q$, define $a_i = |h^{-1}(i)|$. Note that $\sum_{i=1}^q a_i = n$.

We choose independent matrices $B^{(1)}, \dots, B^{(q)}$, with each $B^{(i)}$ as in Theorem 20 with parameters a and v . Here $B^{(i)}$ is a $va \times a_i$ matrix. Finally, let P be an $n \times n$ permutation matrix which, when applied to a matrix A , maps the rows of A in the set $h^{-1}(1)$ to the set of rows $\{1, 2, \dots, a_1\}$, maps the rows of A in the set $h^{-1}(2)$ to the set of rows $\{a_1 + 1, \dots, a_1 + a_2\}$, and for a general $i \in [q]$, maps the set of rows of A in the set $h^{-1}(i)$ to the set of rows $\{a_1 + a_2 + \dots + a_{i-1} + 1, \dots, a_1 + a_2 + \dots + a_i\}$.

The map S is defined to be the product of a block-diagonal matrix and the matrix P :

$$S \equiv \begin{bmatrix} B^{(1)} & & & \\ & B^{(2)} & & \\ & & \ddots & \\ & & & B^{(q)} \end{bmatrix} \cdot P$$

Lemma 22 *$S \cdot A$ can be computed in $O(\text{nnz}(A)(\log(r/\varepsilon\delta))/\varepsilon)$ time.*

Proof: As P is a permutation matrix, $P \cdot A$ can be computed in $O(\text{nnz}(A))$ time and has the same number of non-zero entries of A . For each non-zero entry of $P \cdot A$, we multiply it by $B^{(i)}$ for some i , which takes $O(a) = O(\log(r/\varepsilon\delta)/\varepsilon)$ time. Hence, the total time to compute $S \cdot A$ is $O(\text{nnz}(A)(\log(r/\varepsilon\delta))/\varepsilon)$. ■

5.3 Analysis

We adapt the analysis given for sparse embedding matrices to generalized sparse embedding matrices. Again let $U \in \mathbb{R}^{n \times r}$ have columns that form an orthonormal basis for the column space $C(A)$. Let $U_{1,*}, \dots, U_{n,*}$ be the rows of U , and let $u_i \equiv \|U_{i,*}\|^2$. For $\delta \in (0, 1)$, we set the parameter:

$$T \equiv \frac{r}{C_T q \log(t/\delta)} = \frac{O(\varepsilon^2)}{\log(r/\varepsilon\delta)(r + \log(1/\varepsilon\delta))}, \quad (5)$$

where C_T is a sufficiently large absolute constant.

5.3.1 Vectors with small entries

Let $s \equiv \min\{i' \mid u_i \leq T\}$, and for $y' \in C(A)$ of at most unit norm, let $y \equiv y'_{s:n}$. Since $y_i^2 \leq u_i$, this implies that $\|y\|_\infty^2 \leq T$. Since P is a permutation matrix, we have $\|Py\|_\infty^2 \leq T$.

In this case, we can reduce the analysis to that of a sparse embedding matrix. Indeed, observe that the matrix $B^{(i)} \in \mathbb{R}^{va \times a_i}$ is the concatenation of matrices $\Phi_1^{(i)} D_1^{(i)}, \dots, \Phi_a^{(i)} D_a^{(i)}$, where each $\Phi_j^{(i)} D_j^{(i)} \in \mathbb{R}^{v \times a_i}$ is a sparse embedding matrix. Now fix a value $j \in [a]$ and consider the block-diagonal matrix $N_j \in \mathbb{R}^{qv \times a_i}$:

$$N_j \equiv \begin{bmatrix} \Phi_j^{(1)} D_j^{(1)} & & & \\ & \Phi_j^{(2)} D_j^{(2)} & & \\ & & \ddots & \\ & & & \Phi_j^{(q)} D_j^{(q)} \end{bmatrix} \cdot P$$

Lemma 23 N_j is a random sparse embedding matrix with $qv = t/a$ rows and n columns.

Proof: N_j has a single non-zero entry in each column, and the value of this non-zero entry is random in $\{+1, -1\}$. Hence, it remains to show that the distribution of locations of the non-zero entries of N_j is the same as that in a sparse embedding matrix. This follows from the distribution of the values a_1, \dots, a_q , and the definition of P . ■

Lemma 24 Let $\delta \in (0, 1)$. For $j = 1, \dots, a$, let \mathcal{E}_h^j be the event \mathcal{E}_h of Lemma 2, applied to matrix N_j , with $\delta_h \equiv \delta/a$, and $W \equiv T \log(qv/\delta_h) + r/qv \leq 2r/C_T q$. Suppose $\cap_{j \in [a]} \mathcal{E}_h^j$ holds. This event has probability at least $1 - \delta$. Then there is an absolute constant K_L such that with failure probability at most δ_L ,

$$|\|Sy_{s:n}\|^2 - \|y_{s:n}\|^2| \leq K_L \sqrt{W \log(a/\delta_L)}.$$

Proof: We apply Lemma 3 with N_j the sparse embedding matrix ΦD , and qv , the number of rows of N_j , taking on the role of t in Lemma 2, so that the parameter $W = T \log(qv/\delta_h) + r/qv$ as in the lemma statement. (And since $t = avq$, $qv/\delta_h = t/\delta$, so $W = r/C_T q + r/qv \leq 2r/C_T q$.) Since $\|u_{s:n}\|^2 \leq rT$, it suffices for Lemma 2 if qv is at least $2rT/T^2 \log(t/\delta_h) = 2C_T q$, or $v \geq 2C_T$.

With $\delta_h = \delta/a$, by a union bound $\cap_{j \in [a]} \mathcal{E}_h^j$ occurs with failure probability δ , as claimed.

We have, for given N_j , that with failure probability δ_L/a , $|\|N_j y_{s:n}\|^2 - \|y_{s:n}\|^2| \leq K_L \sqrt{W \log(a/\delta_L)}$. Applying a union bound, and using

$$\|Sy_{s:n}\|_2^2 = \frac{1}{a} \sum_{j=1}^a \|N_j y_{s:n}\|_2^2,$$

the result follows. ■

5.3.2 Vectors with large entries

Again, let $s \equiv \min\{i' \mid u_{i'} \leq T\}$. Since $\sum_i u_i = r$, we have

$$s \leq r/T = C_T q \log(t/\delta).$$

The following is a standard non-weighted balls-and-bins analysis.

Lemma 25 Suppose the previously defined constant $C_t > 0$ is sufficiently large. Let \mathcal{E}_{nw} be the event that $|h^{-1}(i) \cap [s]| \leq C_t \log(r/\varepsilon\delta)$, for all $i \in [q]$. Then $\Pr[\mathcal{E}_{nw}] \geq 1 - \delta/r$.

Proof: For any given $i \in [q]$,

$$\mathbf{E}[|h^{-1}(i) \cap [s]|] = s/q \leq C_T \log(t/\delta) = O(\log(r/\epsilon\delta)).$$

Hence, by a Chernoff bound, for a constant $C_t > 0$,

$$\Pr[|h^{-1}(i) \cap [s]| > C_t \log(r/\epsilon\delta)] \leq e^{-\Theta(\log(r/\epsilon\delta))} = \frac{\delta}{rq},$$

The lemma now follows by a union bound over all $i \in [q]$. ■

Lemma 26 *Assume that \mathcal{E}_{nw} holds. Let \mathcal{E}_s be the event that for all $y \in C(A)$, $\|Sy_{1:(s-1)}\|^2 = (1 \pm \varepsilon/2)\|y_{1:(s-1)}\|^2$. Then $\Pr[\mathcal{E}_s] \geq 1 - \delta/r$.*

Proof: For $i = 1, 2, \dots, q$, let L^i be the at most $C_t \log(r/\epsilon\delta)$ -dimensional subspace which is the restriction of the column space $C(A)$ to coordinates j with $h(j) = i$ and $j < s$. By Corollary 21, for any fixed i , with probability at least $1 - (\delta\varepsilon/r)^{C_{subKN}}$, for all $y \in L^i$, $\|Sy\|^2 = (1 \pm \varepsilon)\|y\|^2$. By a union bound and sufficiently large $C_{subKN} > 0$, this holds for all $i \in [q]$ with probability at least $1 - q(\delta\varepsilon/r)^{C_{subKN}} > 1 - \delta/r$. This condition implies \mathcal{E}_s , since $y_{1:(s-1)}$ can be expressed as $\sum_{i \in [q]} y^{(i)}$, where each $y^{(i)} \in L^i$, and letting $\hat{B}^{(i)}$ denote the va rows of S corresponding to entries from $B^{(i)}$,

$$\begin{aligned} \|Sy_{1:(s-1)}\|^2 &= \sum_{i \in [q]} \|\hat{B}^{(i)} y^{(i)}\|^2 \\ &= \sum_{i \in [q]} (1 \pm \varepsilon) \|y^{(i)}\|^2 \\ &= (1 \pm \varepsilon) \|y_{1:(s-1)}\|^2. \end{aligned}$$

A re-scaling to $\varepsilon/2$ completes the proof. ■

5.4 Putting it all together

Now consider any unit vector y in $C(A)$, and write it as $y_{1:(s-1)} + y_{s:n}$. We seek to bound $\langle Sy_{1:(s-1)}, Sy_{s:n} \rangle$. For notational convenience, define the block-diagonal matrix \tilde{N}_j to be the matrix

$$\tilde{N}_j \equiv \begin{bmatrix} 0 & & & & & & & \\ \cdots & & & & & & & \\ 0 & & & & & & & \\ \Phi_j^{(1)} D_j^{(1)} & & & & & & & \\ 0 & & & & & & & \\ \cdots & & & & & & & \\ 0 & & & & & & & \\ & 0 & & & & & & \\ & \cdots & & & & & & \\ & 0 & & & & & & \\ & \Phi_j^{(2)} D_j^{(2)} & & & & & & \\ & 0 & & & & & & \\ & \cdots & & & & & & \\ & 0 & & & & & & \\ & & \ddots & & & & & \\ & & & 0 & & & & \\ & & & \cdots & & & & \\ & & & 0 & & & & \\ & & & & \ddots & & & \\ & & & & & 0 & & \\ & & & & & \cdots & & \\ & & & & & 0 & & \\ & & & & & \Phi_j^{(q)} D_j^{(q)} & & \\ & & & & & 0 & & \\ & & & & & \cdots & & \\ & & & & & 0 & & \end{bmatrix} \cdot P$$

Then $S = \sqrt{1/a} \cdot \sum_{j=1}^a \tilde{N}_j$. Notice that since the set of non-zero rows of \tilde{N}_j and $\tilde{N}_{j'}$ are disjoint for $j \neq j'$,

$$\begin{aligned} \langle Sy_{1:(s-1)}, Sy_{s:n} \rangle &= \frac{1}{a} \sum_{j=1}^a \langle \tilde{N}_j y_{1:(s-1)}, \tilde{N}_j y_{s:n} \rangle \\ &= \frac{1}{a} \sum_{j=1}^a \langle N_j y_{1:(s-1)}, N_j y_{s:n} \rangle, \end{aligned} \tag{6}$$

where by Lemma 23, each N_j is a sparse embedding matrix with $qv = t/a$ rows and n columns.

Lemma 27 *For W as in Lemma 24, and assuming events $\cap_{j=1}^a \mathcal{E}_h^j$, \mathcal{E}_{nw} , and \mathcal{E}_s , there is absolute constant K_C such that with failure probability δ_C ,*

$$|\langle Sy_{1:(s-1)}, Sy_{s:n} \rangle| \leq K_C \sqrt{W \log(a/\delta_C)}.$$

Proof: We generalize Lemma 6 slightly to bound each summand $\langle N_j y_{1:(s-1)}, N_j y_{s:n} \rangle$.

For a given j , and for each $i \geq s$, let

$$z_m \equiv \sum_{i' \in h_j^{-1}(m), i' < s} y_{i'} D_{i'i'}^{(j)},$$

where h_j is the hash function for $\Phi^{(j)}P$. We have for integer $p \geq 1$ using Khintchine's inequality,

$$\begin{aligned}
& \mathbf{E} [\langle N_j y_{1:(s-1)}, N_j y_{s:n} \rangle^{2p}]^{1/p} \\
&= \mathbf{E} \left[\left(\sum_{i \geq s} y_i D_{ii}^{(j)} z_{h_j(i)} \right)^{2p} \right]^{1/p} \\
&\leq C_p \sum_{i \geq s} y_i^2 z_{h_j(i)}^2 = C_p \sum_{m \in h_j([s-1])} z_m^2 \sum_{\substack{i \in h_j^{-1}(m) \\ i \geq s}} y_i^2 \\
&\leq C_p W V_j,
\end{aligned}$$

where $V_j \equiv \sum_{m \in h_j^{-1}([s-1])} z_m^2$, and $C_p \leq \Gamma(p + 1/2)^{1/p} = O(p)$, and the last inequality uses the assumption that \mathcal{E}_h^j holds. Putting $p = \log(a/\delta_C)$ and applying the Markov inequality, we have for all $j \in [a]$ that

$$\Pr[\langle N_j y_{1:(s-1)}, N_j y_{s:n} \rangle^2 \geq e C_p W V_j] \leq 1 - a \exp(-p) = 1 - \delta_C.$$

Moreover, $\frac{1}{a} \sum_{j \in [a]} V_j = \|S y_{1:(s-1)}\|^2$, which under \mathcal{E}_s is at most $(1 + \varepsilon/2) \|y_{1:(s-1)}\|^2 \leq 1 + \varepsilon/2$. Therefore, with failure probability at most δ_C , we have

$$|\langle S y_{1:(s-1)}, S y_{s:n} \rangle| \leq K_C \sqrt{W \log(a/\delta_C)},$$

for an absolute constant K_C . ■

The following is our main theorem in this section.

Theorem 28 *For given $\delta > 0$, with probability at least $1 - \delta$, for $t = O(r \varepsilon^{-4} \log(r/\varepsilon \delta)(r + \log(1/\varepsilon \delta)))$, S is an embedding matrix for A ; that is, for all $y \in C(A)$, $\|S y\|_2 = (1 \pm \varepsilon) \|y\|_2$. S can be applied to A in $O(\text{nnz}(A) \varepsilon^{-1} \log(r/\delta))$ time.*

Proof: Note that

$$t = avq = O([\varepsilon^{-1} \log(r/\varepsilon \delta)][\varepsilon^{-1}][C_t r \varepsilon^{-2}(r + \log(1/\varepsilon \delta))]),$$

yielding the bound claimed. From Lemma 24, event $\cap_{j \in [a]} \mathcal{E}_h^j$ occurs with failure probability at most δ . From Lemma 25 and 26 the joint occurrence of \mathcal{E}_{nw} and \mathcal{E}_s holds with failure probability at most $2\delta/r \leq \delta$. Given these events, from Lemmas 27 and 24, we have with failure probability at most $\delta_L + \delta_C$ that

$$\begin{aligned}
& | \|S y\|^2 - \|y\|^2 | \\
&= | \|S y_{1:(s-1)}\|^2 - \|y_{1:(s-1)}\|^2 + \|S y_{s:n}\|^2 - \|y_{s:n}\|^2 \\
&\quad + 2 \langle S y_{1:(s-1)}, S y_{s:n} \rangle | \\
&\leq (\varepsilon/2) \|y_{1:(s-1)}\|^2 + K_L \sqrt{W \log(a/\delta_L)} + 2K_C \sqrt{W \log(a/\delta_C)},
\end{aligned}$$

where $W \leq 2r/C_T q$.

Setting $\delta_C = \delta_L = \delta K_{sub}^{-r}$, where K_{sub} is from Lemma 8, and recalling that $a = O(\varepsilon^{-1} \log(r/\varepsilon \delta))$, we have

$$W \log(a/\delta_L) \leq \frac{2r \log(a/\delta_L)}{C_T q} = \frac{2\varepsilon^2 O(r + \log(1/\varepsilon \delta))}{C_T(r + \log(1/\varepsilon \delta))} \leq \varepsilon^2 / C'_T,$$

for absolute constant C'_T . Using Lemma 8, we have that with failure probability at most $\delta + \delta + K_{sub}^r (2\delta K_{sub}^{-r}) \leq 4\delta$, that

$$| \|S y\|^2 - \|y\|^2 | \leq \varepsilon/2 + \sqrt{\varepsilon^2 / C'_T} (K_L + 2K_C) \leq \varepsilon$$

for suitable choice of C'_T . Adjusting δ by a constant factor gives the result. ■

6 Approximating Leverage Scores

Let $A \in \mathbb{R}^{n \times d}$ with rank r . Let $U \in \mathbb{R}^{n \times r}$ be an orthonormal basis for $C(A)$. In [20] it was shown how to obtain a $(1 \pm \varepsilon)$ -approximation u'_i to the leverage score u_i for all $i \in [n]$, for a constant $\varepsilon > 0$, in time $O(nd \log n) + O(d^3 \log n \log d)$. Here we improve the running time of this task as follows. We state the running time for constant ε , though for general ε the running time would be $O(\text{nnz}(A) \log n) + \text{poly}(r\varepsilon^{-1} \log n)$.

Theorem 29 *For any constant $\varepsilon > 0$, there is an algorithm which with probability at least $2/3$, outputs a vector (u'_1, \dots, u'_n) so that for all $i \in [n]$, $u'_i = (1 \pm \varepsilon)u_i$. The running time is*

$$O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n).$$

The success probability can be amplified by independent repetition and taking the coordinate-wise median of the vectors u' across the repetitions.

Proof: We first run the algorithm of Theorem 2.6 and Theorem 2.7 of [8]. The first theorem gives an algorithm which outputs the rank r of A , while the second theorem gives an algorithm which also outputs the indices i_1, \dots, i_r of linearly independent columns of A . The algorithm takes $O(\text{nnz}(A) \log d) + O(r^3)$ time and succeeds with probability at least $1 - O(\log d)/d^{1/3}$. Hence, in what follows, we can assume that A has full rank.

We follow the same procedure as Algorithm 1 in [20], using our improved subspace embedding. The proof of [20] proceeds by choosing a subspace embedding Π_1 , computing $\Pi_1 A$, then computing a change of basis matrix R so that $\Pi_1 A R$ has orthonormal columns. The analysis there then shows that the row norms $\|(AR)_{i,*}\|_2^2$ are equal to $u_i(1 \pm \varepsilon)$. To obtain these row norms quickly, an $r \times O(\log n)$ Johnson-Lindenstrauss matrix Π_2 is sampled, and one first computes $R\Pi_2$, followed by $A(R\Pi_2)$. Using a fast Johnson-Lindenstrauss transform Π_1 , one can compute $\Pi_1 A$ in $O(nr \log n)$ time. Π_1 has $O(r \log n \log r)$ rows, and one can compute the $r \times r$ matrix R in $O(r^3 \log n \log r)$ time by computing a QR-factorization. Computing $R\Pi_2$ can be done in $O(r^2 \log n)$ time, and computing $A(R\Pi_2)$ can be done in $O(\text{nnz}(A) \log n)$ time.

Our only change to this procedure is to use a different matrix Π_1 , which is the composition of our subspace embedding matrix S of Theorem 28 with parameter $t = O(r^2 \log r)$, together with a fast Johnson Lindenstrauss transform F . That is, we set $\Pi_1 = F \cdot S$. Here, F is an $O(r \log^2 r) \times t$ matrix, see Section 2.3 of [20] for an instantiation of F . Then, $S \cdot A$ can be computed in $O(\text{nnz}(A) \log r)$ time by Lemma 22. Moreover, $F \cdot (SA)$ can be computed in $O(t \cdot r \log r) = O(r^3 \log^2 r)$ time. One can then compute the matrix R above in $O(r^3 \log^2 r)$ time by computing a QR-factorization of FSA . Then one can compute $R\Pi_2$ in $O(r^2 \log n)$ time, and computing $A(R\Pi_2)$ can be done in $O(\text{nnz}(A) \log n)$ time. Hence, the total time is $O(\text{nnz}(A) \log n + r^3 \log^2 r + r^2 \log n)$ time.

Notice that by Theorem 28, with probability at least $4/5$, $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$ for all $y \in C(A)$, and by Lemma 3 of [20], with probability at least $9/10$, $\|FSy\|_2 = (1 \pm \varepsilon)\|Sy\|_2$ for all $y \in C(A)$. Hence, $\|FSAx\|_2 = (1 \pm \varepsilon)^2 \|Ax\|_2$ for all $x \in \mathbb{R}^d$ with probability at least $7/10$. There is also a small $1/n$ probability of failure that $\|(AR\Pi_2)_{i,*}\|_2 \neq (1 \pm \varepsilon)\|(AR)_{i,*}\|_2$ for some value of i . Hence, the overall success probability is at least $2/3$.

The rest of the correctness proof is identical to the analysis in [20]. ■

7 Least Squares Regression

Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be a matrix and vector for the regression problem: $\min_x \|Ax - b\|_2$. We assume $n > d$. Again, let r be the rank of A . We show that with probability at least $2/3$, we can find an x' for which

$$\|Ax' - b\|_2 \leq (1 + \varepsilon) \min_x \|Ax - b\|_2.$$

We will give several different algorithms. First, we give an algorithm showing that the dependence on $\text{nnz}(A)$ can be linear. Next we shift to the generalized case, with multiple right-hand-sides, and after some analytical preliminaries, give an algorithm based on sampling using leverage scores. Finally, we discuss affine embeddings, constrained regression, and iterative methods.

Theorem 30 *The ℓ_2 -regression problem can be solved up to a $(1 + \varepsilon)$ -factor with probability at least $2/3$ in $O(\text{nnz}(A) + O(d^3 \varepsilon^{-2} \log^7(d/\varepsilon)))$ time.*

Proof: By Theorem 11 applied to the column space $C(A \circ b)$, where $A \circ b$ is A adjoined with the vector b , it suffices to compute $\Phi D A$ and $\Phi D b$ and output $\text{argmin}_x \|\Phi D A x - \Phi D b\|_2$. We use the fact that $d \geq r$, and apply Theorem 19 with $t = O(d^2 \varepsilon^{-2} \log^6(d/\varepsilon))$.

The theorem implies that with probability at least $9/10$, all vectors y in the space spanned by the columns of A and b have their norms preserved up to a $(1 + \varepsilon)$ -factor. Notice that $\Phi D A$ and $\Phi D b$ can be computed in $O(\text{nnz}(A))$ time. Now we have a regression problem with $d' = O(d^2 \varepsilon^{-2} \log^6(d/\varepsilon))$ rows and d columns. Using the Fast Johnson-Lindenstrauss transform, this can be solved in $O(d' d \log(d/\varepsilon) + d^3 \varepsilon^{-1} \log d)$ time, see, Theorem 12 of [50]. The success probability is at least $9/10$. This is $O(d^3 \varepsilon^{-2} \log^7(d/\varepsilon))$ time. \blacksquare

Our remaining algorithms will be stated for generalized regression.

7.1 Generalized Regression and Affine Embeddings

The regression problem can be slightly generalized to

$$\min_X \|AX - B\|_F,$$

where X and B are matrices rather than vectors. This problem, also called *multiple-response* regression, is important in the analysis of our low-rank approximation algorithms, and also of independent interest. Moreover, while an analysis involving the embedding of $A \circ b$ is not significantly different than for an embedding involving A alone, this is not true for $A \circ B$: different techniques must be considered. This subsection gives the needed theorems needed for analyzing algorithms for generalized regression, and also gives a general result for *affine embeddings*.

Another form of sketching matrix relies on sampling based on leverage scores; it will be convenient to define it using sampling with replacement: for given sketching dimension t , for $m \in [t]$ let $S \in \mathbb{R}^{t \times n}$ have $S_{m,z_m} \leftarrow 1/\sqrt{tp_{z_m}}$, where $p_i \geq u_i/2r$, and $z_m = i$ with probability p_i .

The following fact is due to Rudelson[48], but has since seen many proofs, and follows readily from Noncommutative Bernstein inequalities [47], which are very similar to matrix Bernstein inequalities [53].

Fact 31 *For rank- r $A \in \mathbb{R}^{n \times d}$ with row leverage scores u_i , there is $t = O(r \varepsilon^{-2} \log r)$ such that leverage-score sketching matrix $S \in \mathbb{R}^{t \times n}$ is an ε -embedding matrix for A .*

7.2 Preliminaries

We collect a few standard lemmas and facts in this subsection.

Lemma 32 (Approximate Matrix Multiplication) *For A and B matrices with n rows, where A has n columns, and given $\epsilon > 0$, there is $t = \Theta(\epsilon^{-2})$, so that for a $t \times n$ generalized sparse embedding matrix S , or $t \times n$ fast JL matrix, or $t \log(nd) \times n$ subsampled randomized Hadamard matrix, or leverage-score sketching matrix for A under the condition that A has orthonormal columns,*

$$\Pr[\|A^\top S^\top S B - A^\top B\|_F^2 < \epsilon^2 \|A\|_F^2 \|B\|_F^2] \geq 1 - \delta,$$

for any fixed $\delta > 0$.

Proof: For a generalized sparse embedding matrix with parameters k and v , first suppose $v = 1$, so that S is the embedding matrix of §2. Let $X = A^\top S^\top S B - A^\top B$. Then $X_{i,j} = A_i^\top S^\top S B_j - A_i^\top B_j$, where A_i is the i -th column of A and B_j is the j -th column of B . Thorup and Zhang [51] have shown that $\mathbf{E}[X_{i,j}] = 0$ and $\text{Var}[X_{i,j}] = O(1/t) \|A_i\|_2^2 \|B_j\|_2^2$. Consequently, $\mathbf{E}[X_{i,j}^2] = \text{Var}[X_{i,j}] = O(1/t) \cdot \|A_i\|_2^2 \|B_j\|_2^2$, from which for

an appropriate $t = \Theta(\epsilon^{-2})$, the lemma follows by Chebyshev's inequality. For $v > 1$, $X_{i,j} = \frac{v}{t} \sum_{i \in [t/v]} \hat{X}_{i,j}$, see (6), so that

$$\mathbf{Var}[X_{i,j}] = \frac{v^2}{t^2} \sum_i \mathbf{Var}[\hat{X}_{i,j}] \leq \frac{v}{t^2} \|A_i\|_2^2 \|B_j\|_2^2 \leq \frac{1}{t} \|A_i\|_2^2 \|B_j\|_2^2,$$

and similarly the lemma follows for the sparse embedding matrices. The result for fast JL matrices was shown by Sarlós[50], and for subsampled Hadamard by Drineas et al.[26], proof of Lemma 5. (The claim also follows from norm-preserving properties of these transforms, see [31].)

For leverage-score sampling, first note that

$$A^\top S^\top SB - A^\top B = \frac{1}{t} \sum_{\substack{i \in [n] \\ m \in [t]}} A_{i,*}^\top B_{i,*} \left[\frac{\mathbb{I}[z_m = i]}{p_i} - 1 \right]$$

we have $\mathbf{E}[A^\top S^\top SB - A^\top B] = 0$, and using the independence of the z_m , the second moment of $\|A^\top S^\top SB - A^\top B\|_F$ is the expectation of

$$\begin{aligned} & \mathbf{tr}[(A^\top S^\top SB - A^\top B)^\top (A^\top S^\top SB - A^\top B)] \\ &= \frac{1}{t^2} \mathbf{tr} \sum_{\substack{i, i' \in [n] \\ m \in [t]}} B_{i',*}^\top A_{i',*} A_{i,*}^\top B_{i,*} \left[\frac{\mathbb{I}[z_m = i]}{p_i} - 1 \right] \left[\frac{\mathbb{I}[z_m = i']}{p_{i'}} - 1 \right], \end{aligned}$$

which is

$$\frac{1}{t^2} \sum_{m \in [t]} \mathbf{tr} \left[\left[\sum_{i \in [n]} B_{i,*}^\top A_{i,*} A_{i,*}^\top B_{i,*} \frac{1}{p_i} \right] - B^\top A A^\top B \right],$$

or using the cyclic property of the trace, the fact that $p_i \geq \|A_{i,*}\|^2 / 2 \|A\|^2$, and the fact that $\mathbf{tr}[B^\top A A^\top B] = \|A^\top B\|^2 \leq \|A\|^2 \|B\|^2$,

$$\frac{1}{t} \left[\sum_{i \in [n]} \|A_{i,*}\|^2 \|B_{i,*}\|^2 \frac{1}{p_i} - \mathbf{tr}[B^\top A A^\top B] \right] \leq \frac{2}{t} \|A\|^2 \|B\|^2,$$

and so the lemma follows for large enough t in $O(\epsilon^{-2})$, by Chebyshev's inequality. ■

Fact 33 *Given $n \times d$ matrix A of rank $k \leq n^{1/2-\gamma}$ for $\gamma > 0$, and $\epsilon > 0$, an $m \times n$ fast JL matrix Π with $m = \Theta(k/\epsilon^2)$ is a subspace embedding for A with failure probability at most δ , for any fixed $\delta > 0$, and requires $O(nd \log n)$ time to apply to A .*

A similar fact holds for subsampled Hadamard transforms.

Fact 34 (Pythagorean Theorem) *If C and D matrices with the same number of rows and columns, then $C^\top D = 0$ implies $\|C + D\|_F^2 = \|C\|_F^2 + \|D\|_F^2$.*

Fact 35 (Normal Equations) *Given $n \times d$ matrix C , and $n \times d'$ matrix D consider the problem*

$$\min_{X \in \mathbb{R}^{d \times d'}} \|CX - D\|_F^2.$$

The solution to this problem is $X^ = C^- D$, where C^- is the Moore-Penrose inverse of C . Moreover, $C^\top (CX^* - D) = 0$, and so if c is any vector in the column space of C , then $c^\top (CX^* - D) = 0$. Using Fact 34, for any X ,*

$$\|CX - D\|_F^2 = \|C(X - X^*)\|_F^2 + \|CX^* - D\|_F^2.$$

7.3 Generalized Regression: Conditions

The main theorem in this subsection is the following. It could be regarded as a generalization of Lemma 1 of [26].

Theorem 36 *Suppose A and B are matrices with n rows, and A has rank at most r . Suppose S is a $t \times n$ matrix, and the event occurs that S satisfies Lemma 32 with error parameter $\sqrt{\epsilon/r}$, and also that S is a subspace embedding for A with error parameter $\epsilon_0 \leq 1/\sqrt{2}$. Then if \tilde{Y} is the solution to*

$$\min_Y \|S(AY - B)\|_F^2, \quad (7)$$

and Y^ is the solution to*

$$\min_Y \|AY - B\|_F^2, \quad (8)$$

then

$$\|A\tilde{Y} - B\|_F \leq (1 + \epsilon)\|AY^* - B\|_F.$$

Before proving Theorem 36, we will need the following lemma.

Lemma 37 *For S, A, B, Y^* and \tilde{Y} as in Theorem 36, assume that A has orthonormal columns. Then*

$$\|A(\tilde{Y} - Y^*)\|_F \leq 2\sqrt{\epsilon}\|B - AY^*\|_F.$$

Proof: The proof is in the appendix. ■

Proof of Theorem 36: Let A have the thin SVD $A = U\Sigma V^\top$. Since U is a basis for $C(A)$, there are X^* and \tilde{X} so that $S(U\tilde{X} - B) = S(A\tilde{Y} - B)$ and $UX^* - B = AY^* - B$, and therefore $\|U\tilde{X} - B\| \leq (1 + \epsilon)\|UX^* - B\|$ implies the theorem: we can assume without loss of generality that A has orthonormal columns. With this assumption, and using the Pythagorean Theorem (Fact 34) with the normal equations (Fact 35), and then Lemma 37,

$$\begin{aligned} \|A\tilde{Y} - B\|_F^2 &= \|AY^* - B\|_F^2 + \|A(\tilde{Y} - Y^*)\|_F^2 \\ &\leq \|AY^* - B\|_F^2 + 4\epsilon\|AY^* - B\|_F^2 \\ &\leq (1 + 4\epsilon)\|AY^* - B\|_F^2, \end{aligned}$$

and taking square roots and adjusting ϵ by a constant factor completes the proof. ■

7.4 Generalized Regression: Algorithm

Our main algorithm for regression is given in the proof of the following theorem.

Theorem 38 *Given $A \in \mathbb{R}^{n \times d}$ of rank r , and $B \in \mathbb{R}^{n \times d'}$, the regression problem $\min_Y \|AY - B\|_F$ can be solved up to ϵ relative error with probability at least $2/3$, in time*

$$O(\text{nnz}(A) \log n + r^2(r\epsilon^{-1} + rd' + r \log^2 r + d'\epsilon^{-1} + \log n)),$$

and obtaining a coresset of size $O(r(\epsilon^{-1} + \log r))$.

Proof: We estimate the leverage scores of A to relative error $1/2$, using the algorithm of Theorem 29, which has the side effect of finding r independent columns of A , so that we can assume that $d = r$.

If U is a basis for $C(A)$, then for any X there is a Y so that $UX = AY$, and vice versa, so that conditions satisfied by UX are satisfied by AY . That is, we can (and will hereafter) assume that A has r orthonormal columns, when considering products AY .

We construct a leverage-score sketching matrix S for A with $t = O(r/\varepsilon + r \log r)$, so that Lemma 32 is satisfied for error parameter at most $\sqrt{\varepsilon/r}$. With this t , S will also be an ε -embedding matrix with $\varepsilon < 1/\sqrt{2}$, using Lemma 31. These conditions and Theorem 36 imply that the solution \tilde{Y} to $\min_Y \|S(AY - B)\|$ has

$$\|A\tilde{Y} - B\| \leq (1 + \varepsilon) \min_Y \|AY - B\|.$$

The running time is that for computing the leverage scores, plus the time needed for finding \tilde{Y} , which can be done by computing a QR factorization of SA and then computing $R^{-1}Q^\top SB$, which requires $r^3(\varepsilon^{-1} + \log r) + r^2(\varepsilon^{-1} + \log r)d' + r^3d'$, and the cost bound follows. \blacksquare

7.5 Affine Embeddings

We also use *affine embeddings* for which a stronger condition than Theorem 36 is satisfied.

Theorem 39 Suppose A and B are matrices with n rows, and A has rank at most r . Suppose S is a $t \times n$ matrix, and the event occurs that S satisfies Lemma 32 with error parameter ε/\sqrt{r} , and also that S is a subspace embedding for A with error parameter ε . Let X^* be the solution of $\min_X \|AX - B\|$, and $\tilde{B} \equiv AX^* - B$. For all X of appropriate shape,

$$\|S(AX - B)\|^2 - \|S\tilde{B}\|^2 = (1 \pm 2\varepsilon)\|AX - B\|^2 - \|\tilde{B}\|^2,$$

for $\varepsilon \leq 1/2$. So S is an affine embedding with 2ε relative error up to an additive constant. (That is, a weak embedding.) If also $\|S\tilde{B}\|^2 = (1 \pm \varepsilon)\|\tilde{B}\|^2$, then

$$\|S(AX - B)\|^2 = (1 \pm 3\varepsilon)\|AX - B\|^2, \quad (9)$$

and S is a 3ε -affine embedding.

Note that even when only the weaker first statement holds, the sketch still can be used for optimization, since adding a constant to the objective function of an optimization does not change the solution. Note also that

Proof: If U is a basis for $C(A)$, then for any X there is a Y so that $UX = AY$, and vice versa, so that conditions satisfied by UX are satisfied by AY . That is, we can (and will hereafter) assume that A has r orthonormal columns.

Using the fact that $\|W\|^2 = \text{tr } W^\top W$ for any W , the embedding property, the fact that $\|A\| \leq \sqrt{r}$, and the matrix product approximation condition of Lemma 32,

$$\begin{aligned} & \|S(AX - B)\|^2 - \|S\tilde{B}\|^2 \\ &= \|SA(X - X^*) + S(AX^* - B)\|^2 - \|S\tilde{B}\|^2 \\ &= \|SA(X - X^*)\|^2 - 2\text{tr}[(X - X^*)^\top A^\top S^\top S\tilde{B}] \\ &= \|A(X - X^*)\|^2 \\ &\quad \pm \varepsilon(\|A(X - X^*)\|^2 + 2\|X - X^*\|\|\tilde{B}\|). \end{aligned}$$

The normal equations (Fact 35) imply that $\|AX - B\|^2 = \|A(X - X^*)\|^2 + \|\tilde{B}\|^2$, and using the observation that $(a + b)^2 \leq 2(a^2 + b^2)$ for $a, b \in \mathbb{R}$,

$$\begin{aligned} & \|S(AX - B)\|^2 - \|S\tilde{B}\|^2 - (\|AX - B\|^2 - \|\tilde{B}\|^2) \\ &= \pm \varepsilon(\|A(X - X^*)\|^2 + 2\|X - X^*\|\|\tilde{B}\|) \\ &\leq \pm \varepsilon(\|A(X - X^*)\| + \|\tilde{B}\|)^2 \\ &\leq \pm 2\varepsilon(\|A(X - X^*)\|^2 + \|\tilde{B}\|^2) \\ &= \pm 2\varepsilon\|AX - B\|^2, \end{aligned}$$

and the first statement of the theorem follows. When $\|S\tilde{B}\|^2 = (1 \pm \varepsilon)\|\tilde{B}\|^2$, the second statement follows, since then

$$\|S(AX - B)\|^2 = (1 \pm 2\varepsilon)\|AX - B\|^2 \pm \varepsilon\|\tilde{B}\|^2 = (1 \pm 3\varepsilon)\|AX - B\|^2,$$

using $\|\tilde{B}\| \leq \|AX - B\|$ for all X . ■

To apply this theorem to sparse embeddings, we will need the following lemma.

Lemma 40 *Let A be an $n \times d$ matrix. Let $S \in \mathbb{R}^{t \times n}$ be a randomly chosen sparse embedding matrix for an appropriate $t = \Omega(\varepsilon^{-2})$. Then with probability at least $9/10$,*

$$\|SA\|_F^2 = (1 \pm \varepsilon)\|A\|_F^2.$$

Proof: Please see the appendix. ■

Lemma 41 *Let A be an $n \times d$ matrix. Let $S \in \mathbb{R}^{t \times n}$ be an SRHT matrix for an appropriate $t = \Omega(\varepsilon^{-2}(\log n)^2)$. Then with probability at least $9/10$,*

$$\|SA\|_F^2 = (1 \pm \varepsilon)\|A\|_F^2.$$

Proof: Please see the appendix. ■

Theorem 42 *Let A and B be matrices with n rows, and A has rank at most r . The following conditions hold with fixed nonzero probability. If S is a $t \times n$ sampled randomized Hadamard transform (SRHT) matrix, there is $t = O(\varepsilon^{-2}[\log^2 n + (\log r)(\sqrt{r} + \sqrt{\log n})^2])$ such that S is an ε -affine embedding for A and B . If S is a $t \times n$ sparse embedding, there is $t = O(\varepsilon^{-2}r^2 \log^6(r/\varepsilon))$ such that S is an ε -affine embedding. If S is a $t \times n$ leverage-score sampling matrix, there is $t = O(\varepsilon^{-2}r \log r)$ such that S is a weak ε -affine embedding. If the row norms of \tilde{B} are available, a modified leverage-score sampler is an ε -embedding. (Here \tilde{B} is as in Theorem 39.)*

Note that none of the dimensions t depend on the number of columns of B .

Proof: To apply Theorem 39, we need each given sketching matrix to satisfy conditions on multiplicative error, subspace embedding, and preservation of $\|\tilde{B}\|$. As in that theorem, we can assume without loss of generality that A has r orthonormal columns.

Regarding the multiplicative error bound of ε/\sqrt{r} , Lemma 32 tells us that SRHT achieves this bound for $t = O(\log(n)^2 \varepsilon^{-2} r)$, and the other two need $t = O(\varepsilon^{-2} r)$.

Regarding subspace embedding, as noted in the introduction, an SRHT matrix achieves this for $t = O(\varepsilon^{-2}(\log r)(\sqrt{r} + \sqrt{\log n})^2)$. A sparse embedding requires $t = O(\varepsilon^{-2}r^2 \log^6(r/\varepsilon))$, as in Theorem 19, and leverage score samplers need $t = O(\varepsilon^{-2}r \log r)$, as mentioned in Fact 31.

Regarding preservation of the norm of \tilde{B} , Lemma 41 gives the claim for SRHT matrices, and Lemma 40 gives the claim for sparse embeddings, where the “ A ” of those lemmas is \tilde{B} .

Thus the conditions are satisfied for Theorem 39 to yield the the claims for SRHT and for sparse embeddings, and for the weak condition for leverage score samplers.

We give only a terse version of the argument for the last statement of the theorem. When the squared row norms $b_i \equiv \|\tilde{B}_{i,*}\|^2$ of \tilde{B} are available, a sampler which picks row i with probability $p_i = \min\{1, tb_i/\|\tilde{B}\|^2\}$, and scales that row with $1/\sqrt{tp_i}$, will yield a matrix whose Frobenius norm will be $(1 \pm 1/\sqrt{t})\|\tilde{B}\|$ with high probability. If the leverage score sampler picks rows with probability q_i , create a new sampler that picks rows with probability $p'_i \equiv (p_i + q_i)/2$, and scales by $1/\sqrt{tp'_i}$. The resulting sampler will satisfy the norm preserving property for \tilde{B} , and also satisfy the same properties as the leverage score sampler, up to a constant factor. The resulting sampler is thus an $O(\varepsilon)$ -affine embedding. ■

7.6 Affine Embeddings and Constrained Regression

From the condition (9), an affine embedding can be used to reduce the work needed to achieve small error in regression problems, even when there are constraints on X . We consider the constraint $X \geq 0$, that the entries of X are nonnegative. The problem $\min_{X \geq 0} \|AX - B\|^2$, for $B \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{n \times d}$, arises among other places as a subroutine in finding a nonnegative approximate factorization of B .

For an affine embedding S ,

$$\min_{X \geq 0} \|S(AX - B)\|^2 = (1 \pm \varepsilon) \min_{X \geq 0} \|AX - B\|^2,$$

yielding an immediate reduction yielding a solution with relative error ε : just solve the sketched version of the problem.

From Theorem 42, suitable sketching matrices for constrained regression include a sparse embedding, an SRHT matrix, or a leverage score sampler. (The latter may not need the condition of preserving the norm of \tilde{B} if a high-accuracy solver is used for the sketched solution, or if otherwise the additive constant is not an obstacle for that solver.)

Since it's immediate that affine embeddings can be composed to obtain an affine embedding (with a constant factor loss), the most efficient approach might be use a sketch that first applies a sparse embedding, and then applies an SRHT matrix, resulting in a sketched problem with $O(\varepsilon^{-2} r \log(r/\varepsilon)^2)$ rows, and where computing the sketch takes $O(\text{nnz}(A) + \text{nnz}(B)) + \tilde{O}(\varepsilon^{-2} r^2(d + d'))$ time, for $B \in \mathbb{R}^{n \times d'}$. When r is unknown, the upper bound $r \leq d$ can of course be used.

For low-rank approximation, discussed in §8, we require X to satisfy a rank condition; the same techniques apply.

7.7 Iterative Methods for Regression

A classical approach to finding $\min_X \|AX - B\|$ is to solve the normal equations (Fact 35) $A^\top AX = A^\top B$ via Gaussian elimination; for $A \in \mathbb{R}^{n \times r}$ and $B \in \mathbb{R}^{n \times d'}$, this requires $O(\min\{r \text{nnz}(B), d' \text{nnz}(A)\})$ to form $A^\top B$, $O(r \text{nnz}(A))$ to form $A^\top A$, and $O(r^3 + r^2 d')$ to solve the resulting linear systems. (Another method is to factor $A = QW$, where Q has orthonormal columns and W is upper triangulation; this typically trades a slowdown for a higher-quality solution.)

Another approach to regression is to apply an iterative method (from the general class of Krylov, CG-like methods) to a pre-conditioned version of the problem. In such methods, an estimate $x^{(m)}$ of a solution is maintained, for iterations $m = 0, 1, \dots$, using data obtained from previous iterations. The convergence of these methods depends on the *condition number* $\kappa(A^\top A) = \frac{\sup_{x, \|x\|=1} \|Ax\|^2}{\inf_{x, \|x\|=1} \|Ax\|^2}$ from the input matrix. A classical result ([36] via [40] or Theorem 10.2.6, [28]), is that

$$\frac{\|A(x^{(m)} - x^*)\|^2}{\|A(x^{(0)} - x^*)\|^2} \leq 2 \left(\frac{\sqrt{\kappa(A^\top A)} - 1}{\sqrt{\kappa(A^\top A)} + 1} \right)^m. \quad (10)$$

Thus the running time of CG-like methods, such as CGNR [28], depends on the (unknown) condition number. The running time per iteration is the time needed to compute matrix vector products Ax and $A^\top v$, plus $O(n + d)$ for vector arithmetic, or $O(\text{nnz}(A))$.

Pre-conditioning reduces the number of iterations needed for a given accuracy: suppose for non-singular matrix R , the condition number $\kappa(R^\top A^\top AR)$ is small. Then a CG-like method applied to AR would converge quickly, and moreover for iterate $y^{(m)}$ that has error $\alpha^{(m)} \equiv \|ARy^{(m)} - b\|$ small, the corresponding $x \leftarrow Ry^{(m)}$ would have $\|Ax - b\| = \alpha^{(m)}$. The running time per iteration would have an additional $O(d^2)$ for computing products involving R .

Consider the matrix R obtained for leverage score approximation in §6, where a subspace embedding matrix Π_1 is applied to A , and R is computed so that $\Pi_1 AR$ has orthonormal columns. Since Π_1 is a subspace embedding matrix to constant accuracy ε_0 , for all unit $x \in \mathbb{R}^d$, $\|ARx\|^2 = (1 \pm \varepsilon_0)\|\Pi_1 ARx\|^2 = (1 \pm \varepsilon_0)^2$.

It follows that the condition number

$$\kappa(R^\top A^\top AR) \leq \frac{(1 + \varepsilon_0)^2}{(1 - \varepsilon_0)^2}.$$

That is, AR is very well-conditioned. Plugging this bound into (10), after m iterations $\|AR(x^{(m)} - x^*)\|^2$ is at most $2\varepsilon_0^m$ times its starting value.

Thus starting with a solution $x^{(0)}$ with relative error at most 1, and applying $1 + \log(1/\varepsilon)$ iterations of a CG-like method with $\varepsilon_0 = 1/e$, the relative error is reduced to ε and the work is $O((\text{nnz}(A) + r^2) \log(1/\varepsilon))$ (where we assume d has been reduced to r , as in the leverage computation), plus the work to find R . We have

Theorem 43 *The ℓ_2 -regression problem can be solved up to a $(1 + \varepsilon)$ -factor with probability at least $2/3$ in*

$$O(\text{nnz}(A) \log(n/\varepsilon) + r^3 \log^2 r + r^2 \log(1/\varepsilon))$$

time.

Note that only the matrix R from the leverage score computation is needed, not the leverage scores, so the $\text{nnz}(A)$ term in the running time need not have a $\log(n)$ factor; however, since reducing A to r columns requires that factor, the resulting running time without that factor is $O(\text{nnz}(A) \log(1/\varepsilon) + d^3 \log^2 d + d^2 \log(1/\varepsilon))$, depends on d .

The matrix AR is so well-conditioned that a simple iterative improvement scheme has the same running time up to a constant factor. Again start with a solution $x^{(0)}$ with relative error at most 1, and for $m \geq 0$, let $x^{(m+1)} \leftarrow x^{(m)} + R^\top A^\top (b - ARx^{(m)})$. Then using the normal equations,

$$\begin{aligned} AR(x^{(m+1)} - x^*) &= AR(x^{(m)} + R^\top A^\top (b - ARx^{(m)}) - x^*) \\ &= (AR - ARR^\top A^\top AR)(x^{(m)} - x^*) \\ &= U(\Sigma - \Sigma^3)V^\top(x^{(m)} - x^*), \end{aligned}$$

where $AR = U\Sigma V^\top$ is the SVD of AR .

For all unit $x \in \mathbb{R}^d$, $\|ARx\|^2 = (1 \pm \varepsilon_0)^2$, and so we have that all singular values σ_i of AR are $1 \pm \varepsilon_0$, and the diagonal entries of $\Sigma - \Sigma^3$ are all at most $\sigma_i(1 - (1 - \varepsilon_0)^2) \leq \sigma_i 3\varepsilon_0$ for $\varepsilon_0 \leq 1$. Hence

$$\|AR(x^{(m+1)} - x^*)\| \leq 3\varepsilon_0 \|AR(x^{(m)} - x^*)\|,$$

and by choosing $\varepsilon_0 = 1/2$, say, $O(\log(1/\varepsilon))$ iterations suffice for this scheme also to attain ε relative error.

This scheme can be readily extended to generalized (multiple-response) regression, using the iteration $X^{(m+1)} \leftarrow X^{(m)} + R^\top A^\top (B - ARX^{(m)})$. The initialization cost then includes that of computing $A^\top B$, which is $O(\min\{r \text{nnz}(B), d' \text{nnz}(A)\})$, where again $B \in \mathbb{R}^{n \times d'}$. The product $A^\top A$, used implicitly per iteration, could be computed in $O(r \text{nnz}(A))$, and then applied per iteration in time $d'r^2$, or applied each iteration in time $d' \text{nnz}(A)$.

That is, this method is never much worse than CG-like methods, but comparable in running time when $d' < r$; when $d' > r$, it is a little worse in asymptotic running time than solving the normal equations.

8 Low Rank Approximation

This section gives algorithms for low-rank approximation, understood using generalized regression analysis, as in earlier work such as [50, 10]. Let $\Delta_k \equiv \|A - [A]_k\|_F$, where $[A]_k$ denotes the best rank- k approximation to A . We seek low-rank matrices whose distance to A is within $1 + \varepsilon$ of Δ_k .

While Theorem 11 and Theorem 28 are stated in terms of specific constant probability of success, they can be re-stated and proven so that the failure probabilities are arbitrarily small, but still constant. In the

following we'll assume that adjustments have been done, so that the sum of a fixed number of such failure probabilities is at most $1/5$.

We will apply embedding matrices composed of products of such matrices, so we need to check that this operation preserves the properties we need.

Fact 44 *If $S \in \mathbb{R}^{t \times n}$ approximates matrix products and is a subspace embedding with error ϵ and failure probability δ_S , and $\Pi \in \mathbb{R}^{t \times t}$ approximates matrix products with error ϵ and failure probability δ_Π , then ΠS approximates matrix products with error $O(\epsilon)$ and failure probability at most $\delta_S + \delta_\Pi$.*

Proof: This follows from two applications of Lemma 32, together with the observation that $\|SAx\| = (1 \pm \epsilon)\|Ax\|$ for basis vectors x implies that $\|SA\| = (1 \pm \epsilon)\|A\|$. \blacksquare

Fact 45 *If $S \in \mathbb{R}^{t \times n}$ is a subspace embedding with error ϵ and failure probability δ_S , and $\Pi \in \mathbb{R}^{t \times t}$ is a subspace embedding with error ϵ and failure probability δ_Π , then ΠS is a subspace embedding with error $O(\epsilon)$ and failure probability at most $\delta_S + \delta_\Pi$.*

The following lemma implies a regression algorithm that is linear in $\text{nnz}(A)$, but has a worse dependence in its additive term.

Lemma 46 *Let $A \in \mathbb{R}^{n \times d}$ of rank r , $B \in \mathbb{R}^{n \times d'}$, and $c \equiv d + d'$. For $\hat{R} \in \mathbb{R}^{t \times n}$ a sparse embedding matrix, $\Pi \in \mathbb{R}^{t' \times t}$ a sampled randomized Hadamard matrix, there is $t = O(r^2 \log^6(r/\epsilon) + r\epsilon^{-1})$ and $t' = O(r\epsilon^{-1} \log(r/\epsilon))$ such that for $R \equiv \Pi \hat{R}$, $\tilde{X} \equiv \arg\min_X \|R(AX - B)\|$ has $\|A\tilde{X} - B\| \leq (1 + \epsilon) \min_X \|AX - B\|$. The operator R can be applied in $O(\text{nnz}(A) + \text{nnz}(B) + tc \log t)$ time.*

Theorem 47 *For $A \in \mathbb{R}^{n \times n}$, there is an algorithm that with failure probability $1/10$ finds matrices $L, W \in \mathbb{R}^{n \times k}$ with orthonormal columns, and diagonal $D \in \mathbb{R}^{k \times k}$, so that $\|A - LDW^\top\| \leq (1 + \epsilon)\Delta_k$. The algorithm runs in time*

$$O(\text{nnz}(A)) + \tilde{O}(nk^2\epsilon^{-4} + k^3\epsilon^{-5}).$$

Proof: The algorithm is as follows:

1. Compute AR^\top and an orthonormal basis U for $C(AR^\top)$, where R is as in Lemma 46 with $r = k$;
2. Compute SU and SA for S the product of a $v' \times v$ SRHT matrix with a $v \times n$ sparse embedding, where $v = \Theta(\epsilon^{-4}k^2 \log^6(k/\epsilon))$ and $v' = \Theta(\epsilon^{-3}k \log^2(k/\epsilon))$. (Instead of this affine embedding construction, an alternative might use leverage score sampling, where even the weaker claim of Theorem 42 would be enough.)
3. Compute the SVD of $SU = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$;
4. Compute the SVD $\hat{U}DW^\top$ of $\tilde{V}\tilde{\Sigma}^-[\tilde{U}^\top SA]_k$, where again $[Z]_k$ denotes the best rank- k approximation to matrix Z ;
5. Return $L = U\hat{U}$, D , and W .

Running time. Computing AR^\top in the first step takes $O(\text{nnz}(A) + \tilde{O}(nk(k + \epsilon^{-1})))$ time, and then $\tilde{O}(n(k/\epsilon)^2)$ to compute the $n \times O(\epsilon^{-1}k \log(k/\epsilon))$ matrix U . Computing SU and SA requires $O(\text{nnz}(A)) + \tilde{O}(nk^2\epsilon^{-4})$ time. Computing the SVD of the $\tilde{O}(k\epsilon^{-3}) \times \tilde{O}(k\epsilon^{-1})$ matrix SU requires $\tilde{O}(k^3\epsilon^{-5})$. Computing $\tilde{U}^\top SA$ requires $\tilde{O}(nk^2\epsilon^{-4})$ time. Computing the SVD of the $\tilde{O}(k\epsilon^{-1}) \times n$ matrix of the next step requires $\tilde{O}(nk^2\epsilon^{-2})$ time, as does computing $U\hat{U}$.

Correctness Apply Lemma 46 with A of that lemma mapping to A_k^\top and B mapping to A^\top . Taking transposes, this implies that with small fixed failure probability, $\tilde{Y} \equiv AR^\top(A_k R^\top)^-$ has

$$\|\tilde{Y}A_k - A\| \leq (1 + \epsilon) \min_Y \|YA_k - A\| = (1 + \epsilon)\Delta_k,$$

and so

$$\begin{aligned} \min_{X, \text{rank } X=k} \|AR^\top X - A\| &\leq \|AR^\top(A_k R^\top)^- A_k - A\| \\ &\leq (1 + \epsilon)\Delta_k. \end{aligned} \tag{11}$$

Since U is a basis for $C(AR^\top)$,

$$(1 + \epsilon) \min_{X, \text{rank } X=k} \|UX - A\| \leq (1 + \epsilon) \min_{X, \text{rank } X=k} \|AR^\top X - A\|.$$

With the given construction of S , Theorem 42 applies (twice), with AR^\top taking the role of A , and A taking the role of B , so that S is an ϵ -affine embedding, after adjusting constants. It follows that for $\tilde{X} \equiv \text{argmin}_{X, \text{rank } X=k} \|S(UX - A)\|$,

$$\begin{aligned} \|U\tilde{X} - A\| &\leq (1 + \epsilon) \min_{X, \text{rank } X=k} \|UX - A\| \\ &\leq (1 + \epsilon) \min_{X, \text{rank } X=k} \|AR^\top X - A\| \\ &\leq (1 + \epsilon)^2 \Delta_k, \end{aligned}$$

using (11). From lemma 4.3 of [10], the solution to

$$\min_{X, \text{rank } X=k} \|\tilde{U}X - SA\|$$

is $\hat{X} = [\tilde{U}^\top SA]_k$, where this denotes the best rank- k approximation to $\tilde{U}^\top SA$. It follows that $\tilde{X} = \tilde{V}\tilde{\Sigma}^-\hat{X}$ is a solution to $\min_{X, \text{rank } X=k} \|S(UX - A)\|$. Moreover, the rank- k matrix $U\tilde{X} = LDW^\top$ has $\|LDW^\top - A\| \leq (1 + \epsilon)^2 \Delta_k$, and L , D , and W have the properties promised. \blacksquare

9 ℓ_p -Regression for any $1 \leq p < \infty$

Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be a matrix and vector for the regression problem: $\min_x \|Ax - b\|_p$. We assume $n > d$. Let r be the rank of A . We show that with probability at least $2/3$, we can quickly find an x' for which

$$\|Ax' - b\|_p \leq (1 + \epsilon) \min_x \|Ax - b\|_p.$$

Here p is any constant in $[1, \infty)$.

This theorem is an immediate corollary of Theorem 28 and the construction given in section 3.2 of [9], which shows how to solve ℓ_p -regression given a subspace embedding (for ℓ_2) as a black box. We review the construction of [9] below for completeness.

As in the proof of Theorem 29, in $O(\text{nnz}(A) \log d) + O(r^3)$ time we can replace the input matrix A with a new matrix with the same column space of A and full column rank, where r is rank of A . We therefore assume A has full rank in what follows.

Let $w = \Theta(r^6 \log n(r + \log n))$ and assume $w \mid n$. Split A into n/w matrices $A_1, \dots, A_{n/w}$, each $w \times r$, so that A_i is the submatrix of A indexed by the i -th block of w rows.

We invoke Theorem 28 with the parameters $n = w$, r , $\epsilon = 1/2$, and $\delta = 1/(100n)$, choosing a generalized sparse embedding matrix S with $t = O(r \log n(r + \log n))$ rows. Theorem 28 has the guarantee that for each fixed i , SA_i is a subspace embedding with probability at least $1 - \delta$. It follows by a union bound

that with probability at least $1 - 1/(100w)$, for all $i \in [n/w]$, SA_i is a subspace embedding. We condition on this event occurring.

Consider the matrix $F \in \mathbb{R}^{nt/w \times n}$, which is a block-diagonal matrix comprising n/w blocks along the diagonal. Each block is the $t \times w$ matrix S given above.

$$F \equiv \begin{bmatrix} S & & & \\ & S & & \\ & & \ddots & \\ & & & S \end{bmatrix}$$

We will need the following theorem.

Theorem 48 (Theorem 5 of [11], restated) *Let A be an $n \times r$ matrix, and let $p \in [1, \infty)$. Then there exists an (α, β, p) -well-conditioned basis for the column space of A such that if $p < 2$, then $\alpha = r^{1/2+1/p}$ and $\beta = 1$; if $p = 2$, then $\alpha = r^{1/2}$ and $\beta = 1$, and if $p > 2$ then $\alpha = r^{1/2+1/p}$ and $\beta = r^{1/2-1/p}$. An $r \times r$ change of basis matrix U for which $A \cdot U$ is a well-conditioned basis can be computed in $O(nr^5 \log n)$ time.*

The specific conditions satisfied by a well-conditioned basis are given (and used) in the proof of the theorem below. We use the following algorithm **Condition**(A) given a matrix $A \in \mathbb{R}^{n \times r}$:

1. Compute FA ;
2. Apply Theorem 48 to FA to obtain an $r \times r$ change of basis matrix U so that FAU is an (α, β, p) -well-conditioned basis of the column space of matrix FA ;
3. Output $AU/(r\gamma_p)$, where $\gamma_p \equiv \sqrt{2}t^{1/p-1/2}$ for $p \leq 2$, and $\gamma_p \equiv \sqrt{2}w^{1/2-1/p}$ for $p \geq 2$.

The following lemma is the analogue of that in [9] proved for the Fast Johnson Lindenstrauss Transform. However, the proof in [9] only used that the Fast Johnson Lindenstrauss Transform is a subspace embedding. We state it here with our new parameters, and give the analogous proof in the Appendix for completeness.

Lemma 49 *With probability at least $1 - 1/(100w)$, the output $AU/(r\gamma_p)$ of **Condition**(A) is guaranteed to be a basis that is $(\alpha, \beta\sqrt{3}r(tw)^{|1/p-1/2|}, p)$ -well-conditioned, that is, an $(\alpha, \beta \cdot \text{poly}(\max(r, \log n)), p)$ -well-conditioned basis. The time to compute U is $O(\text{nnz}(A) \log n) + \text{poly}(r\epsilon^{-1})$.*

The following text is from [9], we state it here for completeness. A well-conditioned basis can be used to solve ℓ_p regression problems, via an algorithm based on sampling the rows of A with probabilities proportional to the norms of the rows of the corresponding well-conditioned basis. This entails using for speed a second projection Π_2 applied to AU on the right to estimate the row norms, where Π_2 can be an $O(r) \times O(\log n)$ matrix of i.i.d. normal random variables, which is the same as is done in [20]. This allows fast estimation of the ℓ_2 norms of the rows of AU ; however, we need the ℓ_p norms of those rows, which we thus know up to a factor of $r^{|1/2-1/p|}$. We use these norm estimates in the sampling algorithm of [11]; as discussed for the running time bound of that paper, Theorem 7, this algorithm samples a number of rows proportional to $r(\alpha\beta)^p$, when an (α, β, p) -well-conditioned basis is available. This factor, together with a sample complexity increase of $r^{p|1/2-1/p|} = r^{|p/2-1|}$ needed to compensate for error due to using Π_2 , gives a sample complexity increase for our algorithm over that of [11] of a factor of

$$[r^{|p/2-1|}]r^{p+1}(tw)^{|p/2-1|} = \max(r, \log n)^{O(p)},$$

while the leading term in the complexity (for $n \gg r$) is reduced from $O(nr^5 \log n)$ to $O(\text{nnz}(A) \log n)$.

Observe that if $r < \log n$, then $\text{poly}(r\epsilon^{-1} \log n)$ is less than $n \log n$, which is $O(\text{nnz}(A) \log n)$. Hence, the overall time complexity is $O(\text{nnz}(A) \log n) + \text{poly}(r\epsilon^{-1})$.

We adjust Theorem 4.1 of [11] and obtain the following.

Theorem 50 *Given $\epsilon \in (0, 1)$, a constant $p \in [1, \infty)$, $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, there is a sampling algorithm for ℓ_p regression that constructs a coreset specified by a diagonal sampling matrix D , and a solution vector $\hat{x} \in \mathbb{R}^d$ that minimizes the weighted regression objective $\|D(Ax - b)\|_p$. The solution \hat{x} satisfies, with probability at least $1/2$, the relative error bound that $\|A\hat{x} - b\|_p \leq (1 + \epsilon)\|Ax - b\|_p$ for all $x \in \mathbb{R}^d$. Further, with probability $1 - o(1)$, the entire algorithm to construct \hat{x} runs in time*

$$O(\text{nnz}(A) \log n) + \text{poly}(r\epsilon^{-1}).$$

10 Preliminary Experiments

Some preliminary experiments show that a low-rank approximation technique that is a simplified version of these algorithms is promising, and in practice may perform much better than the general bounds of our results.

Here we apply the algorithm of Theorem 47, except that we skip the randomized Hadamard and simply use a sparse embedding \hat{R} and leverage score sampling. We compare the Frobenius error of the resulting LDW^\top with that of the best rank- k approximation.

In our experiments, the matrices tested are $n \times d$.

The resulting low-rank approximation was tested for t_R (the number of columns of \hat{R}) taking values of the form $\lfloor 1.6^z - 0.5 \rfloor$, for integer $z \geq 1$, while $t_R \leq d/5$. The number t_S of rows of S was chosen such that the condition number of SU was at most 1.2. (Since U has orthogonal columns, its condition number is 1, so a large enough leverage score sample will have this property.) For such t_R and t_S , we took the ratio R_e of the Frobenius norm of the error to the Frobenius norm of the error of the best rank- k approximation. The resulting points $(k/t_R, R_e - 1)$ were generated, for all test matrices, for three independent trials, resulting in a set of points P .

The test matrices are from the University of Florida Sparse Matrix Collection, essentially most of those with at most 10^5 nonzero entries, and with n up to about 7000. There were 1155 matrices tested, from 70 sub-collections of matrices, each such sub-collection representing a particular application area.

The curve in Figure 1 represents the results of these tests, where for a particular point (x, y) on the curve, at most one percent of points $(t/k_R, R_e - 1) \in P$ gave a result where $k/t_R < x$ but $R_e - 1 > y$.

Figure 2 shows a similar curve for the points $(t_R/t_S, \text{cond}(SU) - 1)$; thus the necessary ratio t_R/t_S , so that $\text{cond}(SU) \leq 1.2$, as for the results in Figure 1, need be no smaller than about $1/110$.

Acknowledgements

We acknowledge the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. We thank Jelani Nelson and the anonymous STOC referees for helpful comments.

References

- [1] Dimitris Achlioptas, Amos Fiat, Anna R. Karlin, and Frank McSherry. Web search via hub synthesis. In *FOCS*, pages 500–509, 2001.
- [2] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *COLT*, pages 458–469, 2005.
- [3] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2), 2007.
- [4] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *APPROX-RANDOM*, pages 272–279, 2006.

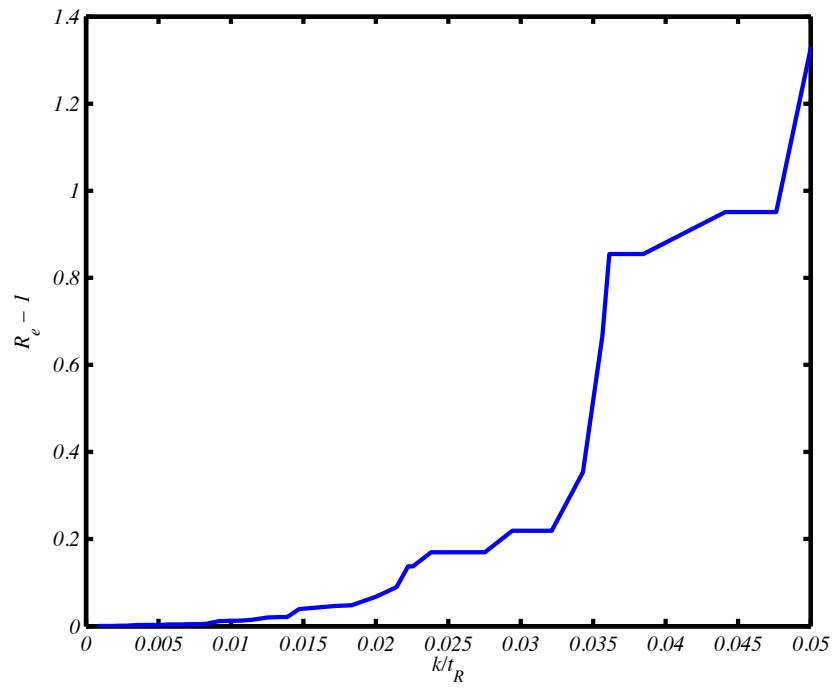


Figure 1: A “1%-Pareto” curve of error as a function of the size of \hat{R}

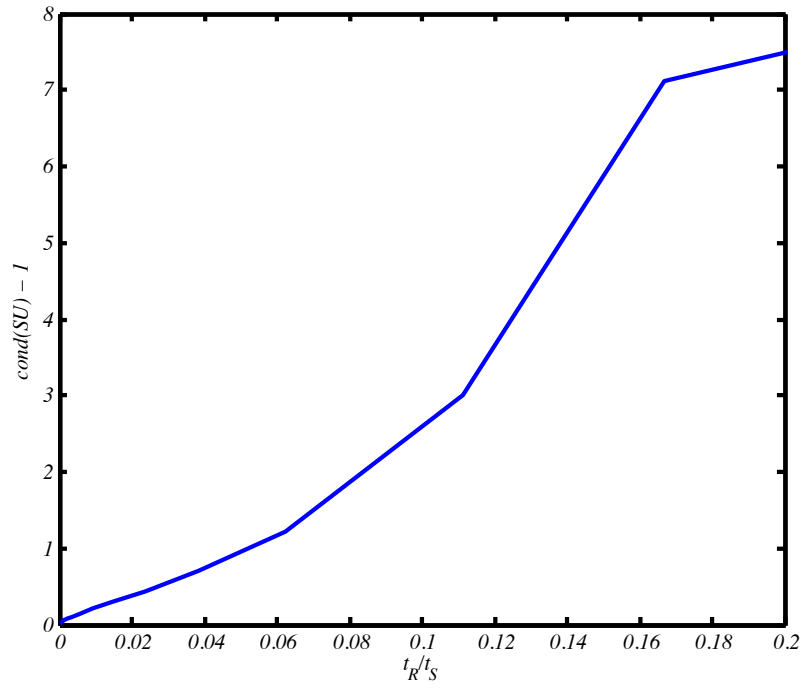


Figure 2: A 1%-Pareto curve of $\text{cond}(SU) - 1$ as a function of the size of \hat{S} relative to \hat{R}

- [5] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *STOC*, pages 619–626, 2001.
- [6] C. Boutsidis and A. Gittens. Improved matrix algorithms via the Subsampled Randomized Hadamard Transform. *ArXiv e-prints*, March 2012.
- [7] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.
- [8] Ho Yee Cheung, Tsz Chiu Kwok, and Lap Chi Lau. Fast matrix rank algorithms and applications. In *STOC*, pages 549–562, 2012.
- [9] K. Clarkson, P. Drineas, Malik Magdon-Ismail, M. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast Cauchy transform and faster robust linear regression. In *SODA*, 2013.
- [10] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.
- [11] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.
- [12] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *STOC*, pages 341–350, 2010.
- [13] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *SODA*, pages 1117–1126, 2006.
- [14] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *APPROX-RANDOM*, pages 292–303, 2006.
- [15] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.
- [16] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.
- [17] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006.
- [18] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1):184–206, 2006.
- [19] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *STOC*, pages 82–90, 2002.
- [20] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *CoRR*, abs/1109.3843, 2011.
- [21] Petros Drineas, Michael Mahoney, Malik Magdon-Ismail, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *ICML*, 2012.
- [22] Petros Drineas and Michael W. Mahoney. Approximating a Gram matrix for improved kernel-based learning. In *COLT*, pages 323–337, 2005.
- [23] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *SODA*, pages 1127–1136, 2006.
- [24] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.

- [25] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *ESA*, pages 304–314, 2006.
- [26] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):217–249, 2011.
- [27] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004.
- [28] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.
- [29] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *ArXiv e-prints*, September 2009.
- [30] D.L. Hanson and F.T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.
- [31] Daniel M. Kane and Jelani Nelson. A sparser Johnson-Lindenstrauss transform. *CoRR*, abs/1012.1577, 2010.
- [32] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.
- [33] Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *STOC*, pages 745–754, 2011.
- [34] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.
- [35] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
- [36] D.G. Luenberger and Y. Ye. *Linear and nonlinear programming*, volume 116. Springer Verlag, 2008.
- [37] Avner Magen and Anastasios Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *SODA*, pages 1422–1436, 2011.
- [38] Frank McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.
- [39] X. Meng and M. W. Mahoney. Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. *ArXiv e-prints*, October 2012.
- [40] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A Parallel Iterative Solver for Strongly Over- or Under-Determined Systems. *ArXiv e-prints*, September 2011.
- [41] Gary L. Miller and Richard Peng. Iterative approaches to row sampling. *CoRR*, abs/1211.2713, 2012.
- [42] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. *CoRR*, abs/1211.1002, 2012.
- [43] Jelani Nelson and David P. Woodruff. Fast Manhattan sketches in data streams. In *PODS*, pages 99–110, 2010.
- [44] Nam H. Nguyen, Thong T. Do, and Trac D. Tran. A fast and efficient algorithm for low-rank approximation of a matrix. In *STOC*, pages 215–224, 2009.
- [45] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.*, 61(2):217–235, 2000.

- [46] Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismael, and Petros Drineas. Random projections for support vector machines. *CoRR*, abs/1211.6085, 2012.
- [47] Benjamin Recht. A simpler approach to matrix completion. *CoRR*, abs/0910.0651, 2009.
- [48] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.
- [49] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), 2007.
- [50] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.
- [51] Mikkel Thorup and Yin Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, pages 615–624, 2004.
- [52] Lloyd N. Trefethen and David Bau. *Numerical linear algebra*. SIAM, 1997.
- [53] Anastasios Zouzias. A matrix hyperbolic cosine algorithm and applications. *CoRR*, abs/1103.2793, 2011.
- [54] Anastasios Zouzias and Nikolaos M. Freris. Randomized extended Kaczmarz for solving least-squares. *CoRR*, abs/1205.5770, 2012.

A Deferred proofs

Proof of Lemma 37: Since by assumption A has orthonormal columns, $\|A(\tilde{X} - X^*)\|_F = \|A^\top A(\tilde{X} - X^*)\|_F$, so it suffices to bound the latter, or $\|\beta\|_F$ where $\beta \equiv A^\top A(\tilde{X} - X^*)$. By Fact 35, we have

$$A^\top S^\top S(A\tilde{X} - B) = 0. \quad (12)$$

To bound $\|\beta\|_F$, we bound $\|A^\top S^\top SA\beta\|_F$, and then show that this implies that $\|\beta\|_F$ is small. Using that $AA^\top A = A$ and (12), we have

$$\begin{aligned} A^\top S^\top SA\beta &= A^\top S^\top SAA^\top A(\tilde{X} - X^*) \\ &= A^\top S^\top SA(\tilde{X} - X^*) \\ &= A^\top S^\top SA(\tilde{X} - X^*) + A^\top S^\top S(B - A\tilde{X}) \\ &= A^\top S^\top S(B - AX^*). \end{aligned}$$

Using the hypothesis of the theorem,

$$\|A^\top S^\top SA\beta\|_F = \|A^\top S^\top S(B - AX^*)\|_F \leq \sqrt{\epsilon/r} \|A\|_F \|B - AX^*\|_F \leq \sqrt{\epsilon} \|B - AX^*\|_F.$$

To show that this bound implies that $\|\beta\|_F$ is small, we use the subadditivity of $\|\cdot\|_F$ and the property of any conforming matrices C and D , that $\|CD\|_F \leq \|C\|_2 \|D\|_F$, to obtain

$$\|\beta\|_F \leq \|A^\top S^\top SA\beta\|_F + \|A^\top S^\top SA\beta - \beta\|_F \leq \sqrt{\epsilon} \|B - AX^*\|_F + \|A^\top S^\top SA - I\|_2 \|\beta\|_F.$$

By hypothesis, $\|SAx\|^2 = (1 \pm \epsilon_0)\|x\|^2$ for all x , so that $A^\top S^\top SA - I$ has eigenvalues bounded in magnitude by ϵ_0^2 , which implies singular values with the same bound, so that $\|A^\top S^\top SA - I\|_2 \leq \epsilon_0^2$. Thus $\|\beta\|_F \leq \sqrt{\epsilon} \|B - AX^*\|_F + \epsilon_0^2 \|\beta\|_F$, or

$$\|\beta\|_F \leq \sqrt{\epsilon} \|B - AX^*\|_F / (1 - \epsilon_0^2) \leq 2\sqrt{\epsilon} \|B - AX^*\|_F,$$

since $\epsilon_0^2 \leq 1/2$. This bounds $\|\beta\|_F$, and so proves the lemma. ■

Proof of Lemma 40: Let $S = \Phi D$ with associated hash function $h : [n] \rightarrow [t]$. For A_i denoting the i -th column of A , let $A_i(b)$ denote the column vector whose ℓ -th coordinate is 0 if $h(\ell) \neq b$, and whose ℓ -th coordinate is $A_{\ell,i}$ if $h(\ell) = b$. We use the second moment method to bound $\|SA\|_F^2$. For the expectation,

$$\mathbf{E}_{D,h}[\|SA\|_F^2] = \sum_{i \in [d]} \mathbf{E}_{D,h}[\|SA_i\|_2^2] = \sum_{i \in [d]} \sum_{b \in [t]} \mathbf{E}_{D,h}[(\sum_{\ell|h(\ell)=b} A_{\ell,i} D_{\ell,\ell})^2] = \mathbf{E}_h \left[\sum_{i \in [d]} \sum_{b \in [t]} \|A_i(b)\|_2^2 \right] = \|A\|_F^2. \quad (13)$$

For the second moment,

$$\mathbf{E}_{D,h}[\|SA\|_F^4] = \sum_{i \in [d]} \mathbf{E}_{D,h}[\|SA_i\|_2^4] + \sum_{i \neq j \in [d]} \mathbf{E}_{D,h}[\|SA_i\|_2^2 \cdot \|SA_j\|_2^2]. \quad (14)$$

We handle the first term in (14) as follows:

$$\begin{aligned} \mathbf{E}_{D,h}[\|SA_i\|_2^4] &= \mathbf{E}_h \left[\sum_{b,b' \in [t]} \mathbf{E}_D[(SA_i)_b^2 \cdot (SA_i)_{b'}^2] \right] \\ &= \mathbf{E}_h \left[\sum_{b \in [t]} \mathbf{E}_D[(SA_i)_b^4] + \sum_{b \neq b' \in [t]} \mathbf{E}_D[(SA_i)_b^2] \cdot \mathbf{E}_D[(SA_i)_{b'}^2] \right] \\ &= \mathbf{E}_h \left[\sum_{b \in [t]} \mathbf{E}_D[(\sum_{\ell|h(\ell)=b} A_{\ell,i} D_{\ell,\ell})^4] + \sum_{b \neq b' \in [t]} \mathbf{E}_D[(\sum_{\ell|h(\ell)=b} A_{\ell,i} D_{\ell,\ell})^2] \cdot \mathbf{E}_D[(\sum_{\ell|h(\ell)=b'} A_{\ell,i} D_{\ell,\ell})^2] \right] \\ &\leq \mathbf{E}_h \left[\sum_{b \in [t]} \left(\sum_{\ell|h(\ell)=b} A_{\ell,i}^4 + \binom{4}{2} \sum_{\ell < \ell' | h(\ell)=h(\ell')=b} A_{\ell,i}^2 A_{\ell',i}^2 \right) + \sum_{b \neq b' \in [t]} \|A_i(b)\|_2^2 \cdot \|A_i(b')\|_2^2 \right] \\ &\leq \mathbf{E}_h \left[\|A_i\|_4^4 + \frac{6}{t} \|A_i\|_2^4 + \sum_{b \neq b' \in [t]} \|A_i(b)\|_2^2 \cdot \|A_i(b')\|_2^2 \right] \\ &\leq \mathbf{E}_h \left[\sum_{b \in [t]} \|A_i(b)\|_2^4 \right] + \frac{6}{t} \|A_i\|_2^4 + \mathbf{E}_h \left[\sum_{b \neq b' \in [t]} \|A_i(b)\|_2^2 \cdot \|A_i(b')\|_2^2 \right] \\ &\leq \frac{6}{t} \|A_i\|_2^4 + \|A_i\|_2^4. \end{aligned}$$

For the second term in (14), for $i \neq j \in [d]$,

$$\begin{aligned}
\mathbf{E}_{D,h}[\|SA_i\|_2^2 \cdot \|SA_j\|_2^2] &= \mathbf{E}_{D,h} \left[\sum_{b \in [t]} \left(\sum_{\ell|h(\ell)=b} A_{\ell,i} D_{\ell,\ell} \right)^2 \left(\sum_{\ell'|h(\ell')=b} A_{\ell',j} D_{\ell',\ell'} \right)^2 \right] \\
&\quad + \mathbf{E}_{D,h} \left[\sum_{b \neq b' \in [t]} \left(\sum_{\ell|h(\ell)=b} A_{\ell,i} D_{\ell,\ell} \right)^2 \left(\sum_{\ell'|h(\ell')=b'} A_{\ell',j} D_{\ell',\ell'} \right)^2 \right] \\
&= \mathbf{E}_h \left[\sum_{b \in [t]} \left(\sum_{\ell, \ell'|h(\ell)=h(\ell')=b} A_{\ell,i} A_{\ell',i} D_{\ell,\ell} D_{\ell',\ell'} \right) \left(\sum_{\ell, \ell'|h(\ell)=h(\ell')=b} A_{\ell,j} A_{\ell',j} D_{\ell,\ell} D_{\ell',\ell'} \right) \right] \\
&\quad + \mathbf{E}_h \left[\sum_{b \in [t]} \|A_i(b)\|_2^2 \cdot \|A_j(b)\|_2^2 + \sum_{b \neq b' \in [t]} \|A_i(b)\|_2^2 \cdot \|A_j(b')\|_2^2 \right] \\
&= \|A_i\|_2^2 \cdot \|A_j\|_2^2 + \mathbf{E}_h \left[\sum_{b \in [t]} 4 \sum_{\ell < \ell' | h(\ell)=h(\ell')=b} A_{\ell,i} A_{\ell',i} A_{\ell,j} A_{\ell',j} \right],
\end{aligned}$$

where the constant 4 arises because if we choose indices $\ell < \ell'$ from $\left(\sum_{\ell, \ell' | h(\ell)=h(\ell')=b} A_{\ell,i} A_{\ell',i} D_{\ell,\ell} D_{\ell',\ell'} \right)$ we need to choose the same ℓ and ℓ' from $\left(\sum_{\ell, \ell' | h(\ell)=h(\ell')=b} A_{\ell,j} A_{\ell',j} D_{\ell,\ell} D_{\ell',\ell'} \right)$ in order to have a non-zero expectation, and there are 4 ways of doing this for distinct ℓ, ℓ' . Continuing,

$$\begin{aligned}
\|A_i\|_2^2 \cdot \|A_j\|_2^2 + \mathbf{E}_h \left[\sum_{b \in [t]} 4 \sum_{\ell < \ell' | h(\ell)=h(\ell')=b} A_{\ell,i} A_{\ell',i} A_{\ell,j} A_{\ell',j} \right] &\leq \|A_i\|_2^2 \cdot \|A_j\|_2^2 + \mathbf{E}_h \left[4 \sum_{b \in [t]} \langle A_i(b), A_j(b) \rangle^2 \right] \\
&\leq \|A_i\|_2^2 \cdot \|A_j\|_2^2 + \mathbf{E}_h \left[4 \sum_{b \in [t]} \|A_i(b)\|_2^2 \cdot \|A_j(b)\|_2^2 \right] \\
&= \|A_i\|_2^2 \cdot \|A_j\|_2^2 + \frac{4}{t} \sum_{\ell, \ell' \in [n]} A_{\ell,i}^2 A_{\ell',j}^2 \\
&= \left(1 + \frac{4}{t} \right) \|A_i\|_2^2 \cdot \|A_j\|_2^2.
\end{aligned}$$

Combining (13) with (14) and the bounds on the terms in (14) above,

$$\begin{aligned}
\mathbf{Var}[\|SA\|_F^2] &\leq \left(\sum_{i \in [d]} \frac{6}{t} \|A_i\|_2^4 + \|A_i\|_2^4 \right) + \sum_{i \neq j \in [d]} \left(1 + \frac{4}{t} \right) \|A_i\|_2^2 \cdot \|A_j\|_2^2 - \|A\|_F^2 \\
&\leq \frac{6}{t} \|A\|_F^2 \\
&= \frac{6}{t} \mathbf{E}[\|SA\|_F^2].
\end{aligned}$$

The lemma now follows by Chebyshev's inequality, for appropriate $t = \Omega(\varepsilon^{-2})$. ■

Proof of Lemma 41: Lemma 15 of [6] shows that $\|SA\| \leq (1 + \varepsilon)\|A\|$ with arbitrarily low failure probability, and the other direction follows from a similar argument. Briefly: the expectation of $\|SA\|^2$ is $\|A\|^2$, by construction, and Lemma 11 of [6] implies that with arbitrarily small failure probability, all rows

of SA will have squared norm at most $\beta \equiv \frac{\alpha}{t} \|A\|^2$, where α is a value in $O(\log n)$. Assuming that this bound holds, it follows from Hoeffding's inequality that the probability that $|\|SA\|^2 - \|A\|^2| \geq \varepsilon \|A\|^2$ is at most $2 \exp(-2[\varepsilon \|A\|^2]^2 / t\beta^2)$, or $2 \exp(-2\varepsilon^2 t / \alpha^2)$, so that $t = \Theta(\varepsilon^{-2}(\log n)^2)$ suffices to make the failure probability at most $1/10$. \blacksquare

Proof of Lemma 49: This is almost exactly the same as in [9], we simply adjust notation and parameters. Applying Theorem 28, we have that with probability at least $1 - 1/(100w)$, for all $x \in \mathbb{R}^r$, if we consider $y = Ax$ and write $y^T = [z_1^T, z_2^T, \dots, z_{n/w}^T]$, then for all $i \in [n/w]$,

$$\sqrt{\frac{1}{2}} \|z_i\|_2 \leq \|S z_i\|_2 \leq \sqrt{\frac{3}{2}} \|z_i\|_2$$

By relating the 2-norm and the p -norm, for $1 \leq p \leq 2$, we have

$$\|S z_i\|_p \leq t^{1/p-1/2} \|S z_i\|_2 \leq t^{1/p-1/2} \sqrt{\frac{3}{2}} \|z_i\|_2 \leq t^{1/p-1/2} \sqrt{\frac{3}{2}} \|z_i\|_p,$$

and similarly,

$$\|S z_i\|_p \geq \|S z_i\|_2 \geq \sqrt{\frac{1}{2}} \|z_i\|_2 \geq \sqrt{\frac{1}{2}} w^{1/2-1/p} \|z_i\|_p.$$

If $p > 2$, then

$$\|S z_i\|_p \leq \|S z_i\|_2 \leq \sqrt{\frac{3}{2}} \|z_i\|_2 \leq \sqrt{\frac{3}{2}} w^{1/2-1/p} \|z_i\|_p,$$

and similarly,

$$\|S z_i\|_p \geq t^{1/p-1/2} \|S z_i\|_2 \geq t^{1/p-1/2} \sqrt{\frac{1}{2}} \|z_i\|_2 \geq t^{1/p-1/2} \sqrt{\frac{1}{2}} \|z_i\|_p.$$

Since $\|Ax\|_p^p = \|y\|_p^p = \sum_i \|z_i\|_p^p$ and $\|FAx\|_p^p = \sum_i \|S z_i\|_p^p$, for $p \in [1, 2]$ we have with probability $1 - 1/(100w)$

$$\sqrt{\frac{1}{2}} w^{1/2-1/p} \|Ax\|_p \leq \|FAx\|_p \leq \sqrt{\frac{3}{2}} t^{1/p-1/2} \|Ax\|_p,$$

and for $p \in [2, \infty)$ with probability $1 - 1/(100w)$

$$\sqrt{\frac{1}{2}} t^{1/p-1/2} \|Ax\|_p \leq \|FAx\|_p \leq \sqrt{\frac{3}{2}} w^{1/2-1/p} \|Ax\|_p.$$

In either case,

$$\|Ax\|_p \leq \gamma_p \|FAx\|_p \leq \sqrt{3}(tw)^{|1/p-1/2|} \|Ax\|_p. \quad (15)$$

Applying Theorem 48, we have, from the definition of a (α, β, p) -well-conditioned basis, that

$$\|FAU\|_p \leq \alpha \quad (16)$$

and for all $x \in \mathbb{R}^d$,

$$\|x\|_q \leq \beta \|FAU\|_p. \quad (17)$$

Combining (15) and (16), we have that with probability at least $1 - 1/(100w)$,

$$\|AU/(r\gamma_p)\|_p \leq \sum_i \|AU_i/r\gamma_p\|_p \leq \sum_i \|FAU_i/r\|_p \leq \alpha.$$

Combining (15) and (17), we have that with probability at least $1 - 1/(100w)$, for all $x \in \mathbb{R}^r$,

$$\|x\|_q \leq \beta \|FAUx\|_p \leq \beta \sqrt{3} r (tw)^{|1/p-1/2|} \|AU \frac{1}{r\gamma_p} x\|_p.$$

Hence $AU/(r\gamma_p)$ is an $(\alpha, \beta \sqrt{3} r (tw)^{|1/p-1/2|}, p)$ -well-conditioned basis. The time to compute FA is $O(\mathbf{nnz}(A) \log n)$ by Theorem 28. Notice that FA is an $nt/w \times n$ matrix, which is $O(n/r^5) \times r$, and so the time to compute U from FA is $O((n/r^5)r^5 \log n) = O(\mathbf{nnz}(A) \log n)$, since $\mathbf{nnz}(A) \geq n$. \blacksquare