Fast Regression with an ℓ_∞ Guarantee

Eric Price[1], Zhao Song[2], and David P. Woodruff[3]

1 **Department of Computer Science, The University of Texas at Austin**
ecprice@cs.utexas.edu
2 **Department of Computer Science, The University of Texas at Austin**
zhaos@utexas.edu
3 **IBM Research Almaden**
dpwoodru@us.ibm.com

Abstract

Sketching has emerged as a powerful technique for speeding up problems in numerical linear algebra, such as regression. In the overconstrained regression problem, one is given an $n \times d$ matrix A, with $n \gg d$, as well as an $n \times 1$ vector b, and one wants to find a vector $\hat{x}$ so as to minimize the residual error $\|Ax - b\|_2$. Using the sketch and solve paradigm, one first computes $S \cdot A$ and $S \cdot b$ for a randomly chosen matrix S, then outputs $x' = (SA)^\dagger Sb$ so as to minimize $\|SAx' - Sb\|_2$.

The sketch-and-solve paradigm gives a bound on $\|x' - x^*\|_2$ when A is well-conditioned. Our main result is that, when S is the subsampled randomized Fourier/Hadamard transform, the error $x' - x^*$ behaves as if it lies in a "random" direction within this bound: for any fixed direction $a \in \mathbb{R}^d$, we have with $1 - d^{-c}$ probability that

$$\langle a, x' - x^* \rangle \lesssim \frac{\|a\|_2 \|x' - x^*\|_2}{d^{\frac{1}{2}-\gamma}}, \tag{1}$$

where $c, \gamma > 0$ are arbitrary constants. This implies $\|x' - x^*\|_\infty$ is a factor $d^{\frac{1}{2}-\gamma}$ smaller than $\|x' - x^*\|_2$. It also gives a better bound on the generalization of x' to new examples: if rows of A correspond to examples and columns to features, then our result gives a better bound for the error introduced by sketch-and-solve when classifying fresh examples. We show that not all oblivious subspace embeddings S satisfy these properties. In particular, we give counterexamples showing that matrices based on Count-Sketch or leverage score sampling do not satisfy these properties.

We also provide lower bounds, both on how small $\|x' - x^*\|_2$ can be, and for our new guarantee (1), showing that the subsampled randomized Fourier/Hadamard transform is nearly optimal. Our lower bound on $\|x' - x^*\|_2$ shows that there is an $O(1/\epsilon)$ separation in the dimension of the optimal oblivious subspace embedding required for outputting an x' for which $\|x' - x^*\|_2 \leq \epsilon \|Ax^* - b\|_2 \cdot \|A^\dagger\|_2$, compared to the dimension of the optimal oblivious subspace embedding required for outputting an x' for which $\|Ax' - b\|_2 \leq (1+\epsilon)\|Ax^* - b\|_2$, that is, the former problem requires dimension $\Omega(d/\epsilon^2)$ while the latter problem can be solved with dimension $O(d/\epsilon)$. This explains the reason known upper bounds on the dimensions of these two variants of regression have differed in prior work.

1998 ACM Subject Classification F.2.1 Numerical Algorithms and Problems

Keywords and phrases Linear regression, Count-Sketch, Gaussians, Leverage scores, ℓ_∞-guarantee

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.

1 Introduction

Oblivious subspace embeddings (OSEs) were introduced by Sarlos [23] to solve linear algebra problems more quickly than traditional methods. An OSE is a distribution of matrices

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl;
Article No. ; pp. :1–:14

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

$S \in \mathbb{R}^{m \times n}$ with $m \ll n$ such that, for any d-dimensional subspace $U \subset \mathbb{R}^n$, with "high" probability S preserves the norm of every vector in the subspace. OSEs are a generalization of the classic Johnson-Lindenstrauss lemma from vectors to subspaces. Formally, we require that with probability $1 - \delta$,

$$\|Sx\|_2 = (1 \pm \epsilon)\|x\|_2$$

simultaneously for all $x \in U$, that is, $(1 - \epsilon)\|x\|_2 \leq \|Sx\|_2 \leq (1 + \epsilon)\|x\|_2$.

A major application of OSEs is to regression. The regression problem is, given $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times d}$ for $n \geq d$, to solve for

$$x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2. \tag{2}$$

Because A is a "tall" matrix with more rows than columns, the system is overdetermined and there is likely no solution to $Ax = b$, but regression will find the closest point to b in the space spanned by A. The classic answer to regression is to use the Moore-Penrose pseudoinverse: $x^* = A^\dagger b$ where

$$A^\dagger = (A^\top A)^{-1} A^\top$$

is the "pseudoinverse" of A (assuming A has full column rank, which we will typically do for simplicity). This classic solution takes $O(nd^{\omega-1} + d^\omega)$ time, where $\omega < 2.373$ is the matrix multiplication constant [9, 25, 12]: $nd^{\omega-1}$ time to compute $A^\top A$ and d^ω time to compute the inverse.

OSEs speed up the process by replacing (2) with

$$x' = \arg\min_x \|SAx - Sb\|_2$$

for an OSE S on $d + 1$-dimensional spaces. This replaces the $n \times d$ regression problem with an $m \times d$ problem, which can be solved more quickly since $m \ll n$. Because $Ax - b$ lies in the $d + 1$-dimensional space spanned by b and the columns of A, with high probability S preserves the norm of $SAx - Sb$ to $1 \pm \epsilon$ for all x. Thus,

$$\|Ax' - b\|_2 \leq \frac{1 + \epsilon}{1 - \epsilon} \|Ax^* - b\|_2.$$

That is, S produces a solution x' which preserves the *cost* of the regression problem. The running time for this method depends on (1) the reduced dimension m and (2) the time it takes to multiply S by A. We can compute these for "standard" OSE types:

- If S has i.i.d. Gaussian entries, then $m = O(d/\epsilon^2)$ is sufficient (and in fact, $m \geq d/\epsilon^2$ is required [20]). However, computing SA takes $O(mnd) = O(nd^2/\epsilon^2)$ time, which is worse than solving the original regression problem (one can speed this up using fast matrix multiplication, though it is still worse than solving the original problem).
- If S is a subsampled randomized Hadamard transform (SRHT) matrix with random sign flips (see Theorem 2.4 in [26] for a survey, and also see [8] which gives a recent improvement) then m increases to $\widetilde{O}(d/\epsilon^2 \cdot \log n)$, where $\widetilde{O}(f) = f \operatorname{poly}(\log(f))$. But now, we can compute SA using the fast Hadamard transform in $O(nd \log n)$ time. This makes the overall regression problem take $O(nd \log n + d^\omega/\epsilon^2)$ time.
- If S is a random sparse matrix with random signs (the "Count-Sketch" matrix), then $m = d^{1+\gamma}/\epsilon^2$ suffices for $\gamma > 0$ a decreasing function of the sparsity [5, 18, 19, 3, 6].

(The definition of a Count-Sketch matrix is, for any $s \geq 1$, $S_{i,j} \in \{0, -1/\sqrt{s}, 1/\sqrt{s}\}$, $\forall i \in [m], j \in [n]$ and the column sparsity of matrix S is s. Independently in each column s positions are chosen uniformly at random without replacement, and each chosen position is set to $-1/\sqrt{s}$ with probability 1/2, and $+1/\sqrt{s}$ with probability 1/2.) Sparse OSEs can benefit from the sparsity of A, allowing for a running time of $\widetilde{O}(\text{nnz}(A)) + \widetilde{O}(d^{\omega}/\epsilon^2)$, where $\text{nnz}(A)$ denotes the number of non-zeros in A.

When n is large, the latter two algorithms are substantially faster than the naïve $nd^{\omega-1}$ method.

1.1 Our Contributions

Despite the success of using subspace embeddings to speed up regression, often what practitioners are interested is not in preserving the cost of the regression problem, but rather in the *generalization* or *prediction* error provided by the vector x'. Ideally, we would like for any future (unseen) example $a \in \mathbb{R}^d$, that $\langle a, x' \rangle \approx \langle a, x^* \rangle$ with high probability.

Ultimately one may want to use x' to do classification, such as regularized least squares classification (RLSC) [22], which has been found in cases to do as well as support vector machines but is much simpler [27]. In this application, given a training set of examples with multiple (non-binary) labels identified with the rows of an $n \times d$ matrix A, one creates an $n \times r$ matrix B, each column indicating the presence or absence of one of the r possible labels in each example. One then solves the multiple response regression problem $\min_X \|AX - B\|_F$, and uses X to classify future examples. A commonly used method is for a future example a, to compute $\langle a, x_1 \rangle, \ldots, \langle a, x_r \rangle$, where $x_1, \ldots, x_r$ are the columns of X. One then chooses the label i for which $\langle a, x_i \rangle$ is maximum.

For this to work, we would like the inner products $\langle a, x'_1 \rangle, \ldots, \langle a, x'_r \rangle$ to be close to $\langle a, x^*_1 \rangle, \ldots, \langle a, x^*_r \rangle$, where X' is the solution to $\min_X \|SAX - SB\|_F$ and X^* is the solution to $\min_X \|AX - B\|_F$. For any $O(1)$-accurate OSE on $d + r$ dimensional spaces [23], which also satisfies so-called approximate matrix multiplication with error $\epsilon' = \epsilon/\sqrt{(d+r)}$, we get that

$$\|x' - x^*\|_2 \leq O(\epsilon) \cdot \|Ax^* - b\|_2 \cdot \|A^\dagger\|_2 \tag{3}$$

where $\|A^\dagger\|$ is the spectral norm of $A^\dagger$, which equals the reciprocal of the smallest singular value of A. To obtain a generalization error bound for an unseen example a, one has

$$|\langle a, x^* \rangle - \langle a, x' \rangle| = |\langle a, x^* - x' \rangle| \leq \|x^* - x'\|_2 \|a\|_2 = O(\epsilon)\|a\|_2 \|Ax^* - b\|_2 \|A^\dagger\|_2, \tag{4}$$

which could be tight if given only the guarantee in (3). However, if the difference vector $x' - x^*$ were distributed in a uniformly random direction subject to (3), then one would expect an $\widetilde{O}(\sqrt{d})$ factor improvement in the bound. This is what our main theorem shows:

▶ **Theorem 1** (Main Theorem, informal). *Suppose $n \leq \text{poly}(d)$. Let S be a subsampled randomized Hadamard transform matrix with $m = d^{1+\gamma}/\epsilon^2$ rows for an arbitrarily small constant $\gamma > 0$. For $x' = \arg\min_x \|SAx - Sb\|_2$ and $x^* = \arg\min_x \|Ax - b\|_2$, and any fixed $a \in \mathbb{R}^d$,*

$$|\langle a, x^* \rangle - \langle a, x' \rangle| \leq \frac{\epsilon}{\sqrt{d}} \|a\|_2 \|Ax^* - b\|_2 \|A^\dagger\|_2. \tag{5}$$

with probability $1 - 1/d^C$ for an arbitrarily large constant $C > 0$. This implies that

$$\|x^* - x'\|_\infty \leq \frac{\epsilon}{\sqrt{d}} \|Ax^* - b\|_2 \|A^\dagger\|_2. \tag{6}$$

with $1 - 1/d^{C-1}$ *probability.*

If $n > \text{poly}(d)$*, then by first composing* S *with a Count-Sketch OSE with* $\text{poly}(d)$ *rows, one can achieve the same guarantee.*

(Here γ is a constant going to zero as n increases, see Theorem 10 for a formal statement of Theorem 1).

Notice that Theorem 1 is considerably stronger than that of (4) provided by existing guarantees. Indeed, in order to achieve the guarantee (6) in Theorem 1, one would need to set $\epsilon' = \epsilon/\sqrt{d}$ in existing OSEs, resulting in $\Omega(d^2/\epsilon^2)$ rows. In contrast, we achieve only $d^{1+\gamma}/\epsilon^2$ rows. We can improve the bound in Theorem 1 to $m = d/\epsilon^2$ if S is a matrix of i.i.d. Gaussians; however, as noted, computing $S \cdot A$ is slower in this case.

Note that Theorem 1 also *makes no distributional assumptions* on the data, and thus the data could be heavy-tailed or even adversarially corrupted. This implies that our bound is still useful when the rows of A are not sampled independently from a distribution with bounded variance.

The ℓ_∞ bound (6) of Theorem 1 is achieved by applying (5) to the standard basis vectors $a = e_i$ for each $i \in [d]$ and applying a union bound. This ℓ_∞ guarantee often has a more natural interpretation than the ℓ_2 guarantee—if we think of the regression as attributing the observable as a sum of various factors, (6) says that the contribution of each factor is estimated well. One may also see our contribution as giving a way for estimating the pseudoinverse $A^\dagger$ *entrywise*. Namely, we get that $(SA)^\dagger S \approx A^\dagger$ in the sense that each entry is within additive $O(\epsilon\sqrt{\frac{\log d}{d}}\|A^\dagger\|_2)$. There is a lot of work on computing entries of inverses of a matrix, see, e.g., [1, 16].

Another benefit of the ℓ_∞ guarantee is when the regression vector x^* is expected to be k-*sparse* (e.g. [14]). In such cases, thresholding to the top k entries will yield an ℓ_2 guarantee a factor $\sqrt{k/d}$ better than (3).

One could ask if Theorem 1 also holds for sparse OSEs, such as the Count-Sketch. Surprisingly, we show that one cannot achieve the generalization error guarantee in Theorem 1 with high probability, say, $1 - 1/d$, using such embeddings, despite the fact that such embeddings do approximate the cost of the regression problem up to a $1 + \epsilon$ factor with high probability. This shows that the generalization error guarantee is achieved by some subspace embeddings but not all.

▶ **Theorem 2** (Not all subspace embeddings give the ℓ_∞ guarantee)**.** *The Count-Sketch matrix with* $d^{1.5}$ *rows and sparsity* $d^{.25}$*—which is an OSE with exponentially small failure probability—with constant probability will have a result* x' *that does not satisfy the* ℓ_∞ *guarantee (6).*

We can show that Theorem 1 holds for S based on the Count-Sketch OSE T with $d^{O(C)}/\epsilon^2$ rows with $1 - 1/d^C$ probability. We can thus compose the Count-Sketch OSE with the SRHT matrix and obtain an $O(\text{nnz}(A)) + \text{poly}(d/\epsilon)$ time algorithm to compute $S \cdot TA$ achieving (6). We can also compute $R \cdot S \cdot T \cdot A$, where R is a matrix of Gaussians, which is more efficient now that STA only has $d^{1+\gamma}/\epsilon^2$ rows; this will reduce the number of rows to d/ϵ^2.

Another common method of dimensionality reduction for linear regression is *leverage score sampling* [10, 15, 21, 7], which subsamples the rows of A by choosing each row with probability proportional to its "leverage scores". With $O(d \log(d/\delta)/\epsilon^2)$ rows taken, the result x' will satisfy the ℓ_2 bound (3) with probability $1 - \delta$. However, it does not give a good ℓ_∞ bound:

▶ **Theorem 3** (Leverage score sampling does not give the ℓ_∞ guarantee)**.** *Leverage score sampling with $d^{1.5}$ rows—which satisfies the ℓ_2 bound with exponentially small failure probability—with constant probability will have a result x' that does not satisfy the ℓ_∞ guarantee (6).*

Finally, we show that the $d^{1+\gamma}/\epsilon^2$ rows that SRHT matrices use is roughly optimal:

▶ **Theorem 4** (Lower bounds for ℓ_2 and ℓ_∞ guarantees)**.** *Any sketching matrix distribution over $m \times n$ matrices that satisfies either the ℓ_2 guarantee* (3) *or the ℓ_∞ guarantee* (6) *must have $m \gtrsim \min(n, d/\epsilon^2)$.*

Notice that our result shows the necessity of the $1/\epsilon$ separation between the results originally defined in Equation (3) and (4) of Theorem 12 of [23]. If we want to output some vector x' such that $\|Ax' - b\|_2 \leq (1+\epsilon)\|Ax^* - b\|_2$, then it is known that $m = \Theta(d/\epsilon)$ is necessary and sufficient. However, if we want to output a vector x' such that $\|x' - x^*\|_2 \leq \epsilon\|Ax^* - b\|_2 \cdot \|A^\dagger\|_2$, then we show that $m = \Theta(d/\epsilon^2)$ is necessary and sufficient.

1.1.1 Comparison to Gradient Descent

While this work is primarily about sketching methods, one could instead apply iterative methods such as gradient descent, after appropriately preconditioning the matrix, see, e.g., [2, 28, 5]. That is, one can use an OSE with constant ϵ to construct a preconditioner for A and then run conjugate gradient using the preconditioner. This gives an overall dependence of $\log(1/\epsilon)$.

The main drawback of this approach is that one loses the ability to save on storage space or number of passes when A appears in a stream, or to save on communication or rounds when A is distributed. Given increasingly large data sets, such scenarios are now quite common, see, e.g., [4] for regression algorithms in the data stream model. In situations where the entries of A appear sequentially, for example, a row at a time, one does not need to store the full $n \times d$ matrix A but only the $m \times d$ matrix SA.

Also, iterative methods can be less efficient when solving multiple response regression, where one wants to minimize $\|AX - B\|$ for a $d \times t$ matrix X and an $n \times t$ matrix B. This is the case when ϵ is constant and t is large, which can occur in some applications (though there are also other applications for which ϵ is very small). For example, conjugate gradient with a preconditioner will take $\widetilde{O}(ndt)$ time while using an OSE directly will take only $\widetilde{O}(nd + d^2t)$ time (since one effectively replaces n with O (d) after computing $S \cdot A$), separating t from d. Multiple response regression, arises, for example, in the RLSC application above.

1.1.2 Proof Techniques

Theorem 1. As noted in Theorem 2, there are some OSEs for which our generalization error bound does not hold. This hints that our analysis is non-standard and cannot use generic properties of OSEs as a black box. Indeed, in our analysis, we have to consider matrix products of the form $S^\top S(UU^\top S^\top S)^k$ for our random sketching matrix S and a fixed matrix U, where k is a positive integer. We stress that it is the *same matrix* S appearing multiple times in this expression, which considerably complicates the analysis, and does not allow us to appeal to standard results on approximate matrix product (see, e.g., [26] for a survey). The key idea is to recursively reduce $S^\top S(UU^\top S^\top S)^k$ using a property of S. We use properties that only hold for specifics OSEs S: first, that each column of S is unit vector; and second, that for all pairs (i, j) and $i \neq j$, the inner product between S_i and S_j is at most $\sqrt{\log n}/\sqrt{m}$ with probability $1 - 1/\operatorname{poly}(n)$.

Theorems 2 and 3. To show that Count-Sketch does not give the ℓ_∞ guarantee, we construct a matrix A and vector b as in Figure 1, which has optimal solution x^* with all coordinates $1/\sqrt{d}$. We then show, for our setting of parameters, that there likely exists an index $j \in [d]$ satisfying the following property: the jth column of S has disjoint support from the kth column of S for all $k \in [d+\alpha] \setminus \{j\}$ except for a single $k > d$, for which S_j and S_k share exactly one common entry in their support. In such cases we can compute x'_j explicitly, getting $|x'_j - x^*_j| = \frac{1}{s\sqrt{\alpha}}$. By choosing suitable parameters in our construction, this gives that $\|x' - x^*\|_\infty \gg \frac{1}{\sqrt{d}}$. The lower bound for leverage score sampling follows a similar construction.

Theorem 4 The lower bound proof for the ℓ_2 guarantee uses Yao's minimax principle. We are allowed to fix an $m \times n$ sketching matrix S and design a distribution over $[A\ b]$. We first write the sketching matrix $S = U\Sigma V^\top$ in its singular value decomposition (SVD). We choose the $d+1$ columns of the adjoined matrix $[A, b]$ to be random orthonormal vectors. Consider an $n \times n$ orthonormal matrix R which contains the columns of V as its first m columns, and is completed on its remaining $n-m$ columns to an arbitrary orthonormal basis. Then $S \cdot [A, b] = V^\top RR^\top \cdot [A, b] = [U\Sigma I_m, 0] \cdot [R^\top A, R^\top b]$. Notice that $[R^\top A, R^\top b]$ is equal in distribution to $[A, b]$, since R is fixed and $[A, b]$ is a random matrix with $d+1$ orthonormal columns. Therefore, $S \cdot [A, b]$ is equal in distribution to $[U\Sigma G, U\Sigma h]$ where $[G, h]$ corresponds to the first m rows of an $n \times (d+1)$ uniformly random matrix with orthonormal columns.

A key idea is that if $n = \Omega(\max(m, d)^2)$, then by a result of Jiang [13], any $m \times (d+1)$ submatrix of a random $n \times n$ orthonormal matrix has $o(1)$ total variation distance to a $d \times d$ matrix of i.i.d. $N(0, 1/n)$ random variables, and so any events that would have occurred had G and h been independent i.i.d. Gaussians, occur with the same probability for our distribution up to an $1 - o(1)$ factor, so we can assume G and h are independent i.i.d. Gaussians in the analysis.

The optimal solution x' in the sketch space equals $(SA)^\dagger Sb$, and by using that SA has the form $U\Sigma G$, one can manipulate $\|(SA)^\dagger Sb\|$ to be of the form $\|\tilde{\Sigma}^\dagger(\Sigma R)^\dagger \Sigma h\|_2$, where the SVD of G is $R\tilde{\Sigma}T$. We can upper bound $\|\tilde{\Sigma}\|_2$ by $\sqrt{r/n}$, since it is just the maximum singular value of a Gaussian matrix, where r is the rank of S, which allows us to lower bound $\|\tilde{\Sigma}^\dagger(\Sigma R)^\dagger \Sigma h\|_2$ by $\sqrt{n/r}\|(\Sigma R)^\dagger \Sigma h\|_2$. Then, since h is i.i.d. Gaussian, this quantity concentrates to $\frac{1}{\sqrt{r}}\|(\Sigma R)^\dagger \Sigma h\|$, since $\|Ch\|^2 \approx \|C\|_F^2/n$ for a vector h of i.i.d. $N(0, 1/n)$ random variables. Finally, we can lower bound $\|(\Sigma R)^\dagger \Sigma\|_F^2$ by $\|(\Sigma R)^\dagger \Sigma RR^\top\|_F^2$ by the Pythagorean theorem, and now we have that $(\Sigma R)^\dagger \Sigma R$ is the identity, and so this expression is just equal to the rank of ΣR, which we prove is at least d. Noting that $x^* = 0$ for our instance, putting these bounds together gives $\|x' - x^*\| \geq \sqrt{d/r}$. The last ingredient is a way to ensure that the rank of S is at least d. Here we choose another distribution on inputs A and b for which it is trivial to show the rank of S is at least d with large probability. We require S be good on the mixture. Since S is fixed and good on the mixture, it is good for both distributions individually, which implies we can assume S has rank d in our analysis of the first distribution above.

Notation. For a positive integer, let $[n] = \{1, 2, \ldots, n\}$. For a vector $x \in \mathbb{R}^n$, define $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$ and $\|x\|_\infty = \max_{i \in [n]} |x_i|$. For a matrix $A \in \mathbb{R}^{m \times n}$, define $\|A\|_2 = \sup_x \|Ax\|_2/\|x\|_2$ to be the spectral norm of A and $\|A\|_F = (\sum_{i,j} A_{i,j}^2)^{1/2}$ to be the Frobenius norm of A. We use $A^\dagger$ to denote the Moore-Penrose pseudoinverse of $m \times n$ matrix A, which if $A = U\Sigma V^\top$ is its SVD (where $U \in \mathbb{R}^{m \times n}$, $\Sigma \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{n \times n}$ for $m \geq n$), is given by $A^\dagger = V\Sigma^{-1}U^\top$. In addition to $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq Cg$ (resp. $\geq$) for an absolute constant C. We use $f \eqsim g$ to mean $cf \leq g \leq Cf$ for constants c, C.

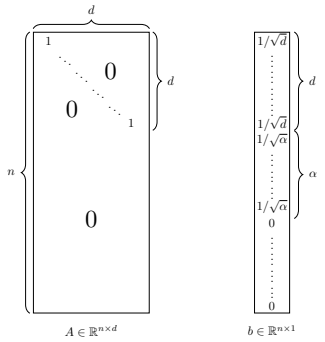

Figure 1 Our construction of A and b for the proof that Count-Sketch does not obey the ℓ_∞ guarantee. $\alpha < d$.

▶ **Definition 5** (Subspace Embedding). A $(1 \pm \epsilon)$ ℓ_2-subspace embedding for the column space of an $n \times d$ matrix A is a matrix S for which for all $x \in \mathbb{R}^d$, $\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2$.

▶ **Definition 6** (Approximate Matrix Product). Let $0 < \epsilon < 1$ be a given approximation parameter. Given matrices A and B, where A and B each have n rows, the goal is to output a matrix C so that $\|A^\top B - C\|_F \leq \epsilon \|A\|_F \|B\|_F$. Typically C has the form $A^\top S^\top SB$, for a random matrix S with a small number of rows. In particular, this guarantee holds for the subsampled randomized Hadamard transform S with $O(\epsilon^{-2})$ rows [11].

Due to space constraints, several proofs are deferred to the full version of our paper.

2 Warmup: Gaussians OSEs

We first show that if S is a Gaussian random matrix, then it satisfies the generalization guarantee. This follows from the rotational invariance of the Gaussian distribution.

▶ **Theorem 7.** *Suppose $A \in \mathbb{R}^{n \times d}$ has full column rank. If the entries of $S \in \mathbb{R}^{m \times n}$ are i.i.d. $N(0, 1/m)$, $m = O(d/\epsilon^2)$, then for any vectors a, b and $x^* = A^\dagger b$, we have, with probability $1 - 1/\operatorname{poly}(d)$,*

$$|a^\top (SA)^\dagger Sb - a^\top x^*| \lesssim \frac{\epsilon\sqrt{\log d}}{\sqrt{d}} \|a\|_2 \|b - Ax^*\|_2 \|A^\dagger\|_2.$$

Because SA has full column rank with probability 1, $(SA)^\dagger SA = I$. Therefore

$$|a^\top (SA)^\dagger Sb - a^\top x^*| = |a^\top (SA)^\dagger S(b - Ax^*)| = |a^\top (SA)^\dagger S(b - AA^\dagger b)|.$$

Thus it suffices to only consider vectors b where $A^\dagger b = 0$, or equivalently $U^\top b = 0$. In such cases, SU will be independent of Sb, which will give the result. The proof is in the full version.

3 SRHT Matrices

We first provide the definition of the subsampled randomized Hadamard transform(SRHT): Let $S = \frac{1}{\sqrt{rn}} PH_n D$, where D is an $n \times n$ diagonal matrix with i.i.d. diagonal entries $D_{i,i}$, for which $D_{i,i}$ in uniform on $\{-1, +1\}$. Here H_n is the Hadamard matrix of size $n \times n$, and we assume n is a power of 2. Here, $H_n = [H_{n/2},\ H_{n/2}; H_{n/2},\ -H_{n/2}]$ and $H_1 = [1]$. The $r \times n$ matrix P samples r coordinates of an n dimensional vector uniformly at random.

For other subspace embeddings, we no longer have that SU and Sb are independent. To analyze them, we start with a claim that allows us to relate the inverse of a matrix to a power series.

▶ Claim 8. Let $S \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{n\times d}$ have SVD $A = U\Sigma V^\top$, and define $T \in \mathbb{R}^{d\times d}$ by $T = I_d - U^\top S^\top SU$. Suppose SA has linearly independent columns and $\|T\|_2 \le 1/2$. Then

$$(SA)^\dagger S = V\Sigma^{-1}\left(\sum_{k=0}^{\infty} T^k\right) U^\top S^\top S. \tag{7}$$

Proof.

$$(SA)^\dagger S = (A^\top S^\top SA)^{-1}A^\top S^\top S = (V\Sigma U^\top S^\top SU\Sigma V^\top)^{-1}V\Sigma U^\top S^\top S$$
$$= V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top S^\top S = V\Sigma^{-1}(I_d - T)^{-1}U^\top S^\top S = V\Sigma^{-1}(\sum_{k=0}^{\infty} T^k)U^\top S^\top S,$$

where in the last equality, since $\|T\|_2 < 1$, the von Neumann series $\sum_{k=0}^{\infty} T^k$ converges to $(I_d - T)^{-1}$. ◀

We then bound the kth term of this sum:

▶ **Lemma 9.** *Let $S \in \mathbb{R}^{r\times n}$ be the subsampled randomized Hadamard transform, and let a be a unit vector. Then with probability $1 - 1/\mathrm{poly}(n)$, we have*

$$|a^\top S^\top S(UU^\top S^\top S)^k b| = O(\log^k n)\cdot(O(d(\log n)/r) + 1)^{\frac{k-1}{2}}\cdot(\sqrt{d}\|b\|_2(\log n)/r + \|b\|_2(\log^{\frac{1}{2}} n)/r^{\frac{1}{2}}).$$

Hence, for r at least $d\log^{2k+2} n\log^2(n/\epsilon)/\epsilon^2)$, this is at most $O(\|b\|_2\epsilon/\sqrt{d})$.

We defer the proof of this lemma to the next section, and now show how the lemma lets us prove that SRHT matrices satisfy the generalization bound with high probability:

▶ **Theorem 10.** *Suppose $A \in \mathbb{R}^{n\times d}$ has full column rank with $\log n = d^{o(1)}$. Let $S \in \mathbb{R}^{m\times n}$ be a subsampled randomized Hadamard transform with $m = O(d^{1+\alpha}/\epsilon^2)$ for $\alpha = \Theta(\sqrt{\frac{\log\log n}{\log d}})$. For any vectors a, b and $x^* = A^\dagger b$, we have*

$$|a^\top(SA)^\dagger Sb - a^\top x^*| \lesssim \frac{\epsilon}{\sqrt{d}}\|a\|_2\|b - Ax^*\|_2\|\Sigma^{-1}\|_2$$

with probability $1 - 1/\mathrm{poly}(d)$.

Proof. Define $\Delta = \Theta\left(\frac{1}{\sqrt{m}}\right)(\log^c d)\|a\|_2\|b - Ax^*\|_2\|\Sigma^{-1}\|_2$. For a constant $c > 0$, we have that S is a $(1\pm\gamma)$ ℓ_2-subspace embedding (Definition 5) for $\gamma = \sqrt{\frac{d\log^c n}{m}}$ with probability $1 - 1/\mathrm{poly}(d)$ (see, e.g., Theorem 2.4 of [26] and references therein), so $\|SUx\|_2 = (1\pm\gamma)\|Ux\|_2$ for all x, which we condition on. Hence for $T = I_d - U^\top S^\top SU$, we have $\|T\|_2 \le (1+\gamma)^2 - 1 \lesssim \gamma$. In particular, $\|T\|_2 < 1/2$ and we can apply Claim 8.

As in Section 2, SA has full column rank if S is a subspace embedding, so $(SA)^\dagger SA = I$ and we may assume $x^* = 0$ without loss of generality.

By the approximate matrix product (Definition 6), we have for some c that

$$|a^\top V\Sigma^{-1}U^\top S^\top Sb| \le \frac{\log^c d}{\sqrt{m}}\|a\|_2\|b\|_2\|\Sigma^{-1}\|_2 \le \Delta, \tag{8}$$

with $1 - 1/\mathrm{poly}(d)$ probability. Suppose this event occurs, bounding the $k = 0$ term of (7). Hence it suffices to show that the $k \ge 1$ terms of (7) are bounded by Δ.

By approximate matrix product, we also have with $1 - 1/d^2$ probability that

$$\|U^\top S^\top Sb\|_F \leq \frac{\log^c d}{\sqrt{m}}\|U^\top\|_F\|b\|_2 \leq \frac{\log^c d\sqrt{d}}{\sqrt{m}}\|b\|_2.$$

Combining with $\|T\|_2 \lesssim \gamma$ we have for any k that

$$|a^\top V\Sigma^{-1}T^kU^\top S^\top Sb| \lesssim \gamma^k(\log^c d)\frac{\sqrt{d}}{\sqrt{m}}\|a\|_2\|\Sigma^{-1}\|_2\|b\|_2.$$

Since this decays exponentially in k at a rate of $\gamma < 1/2$, the sum of all terms greater than k is bounded by the kth term. As long as

$$m \gtrsim \frac{1}{\epsilon^2}d^{1+\frac{1}{k}}\log^c n, \tag{9}$$

we have $\gamma = \sqrt{\frac{d\log^c n}{m}} < \epsilon d^{-1/(2k)}/\log^c n$, so that

$$\sum_{k'\geq k}|a^\top V\Sigma^{-1}T^{k'}U^\top S^\top Sb| \lesssim \frac{\epsilon}{\sqrt{d}}\|a\|_2\|\Sigma^{-1}\|_2\|b\|_2.$$

On the other hand, by Lemma 9 (increasing m by a C^k factor) we have for all k that

$$|a^\top V^\top\Sigma^{-1}U^\top S^\top S(UU^\top S^\top S)^kb| \lesssim \frac{1}{2^k}\frac{\epsilon}{\sqrt{d}}\|a\|_2\|b\|_2\|\Sigma^{-1}\|_2,$$

with probability at least $1 - 1/\operatorname{poly}(d)$, as long as $m \gtrsim d(C\log n)^{2k+2}\log^2(n/\epsilon)/\epsilon^2$, for a sufficiently large constant C. Since the T^k term can be expanded as a sum of 2^k terms of this form, we get that

$$\sum_{k'=1}^{k}|a^\top V\Sigma^{-1}T^{k'}U^\top S^\top Sb| \lesssim \frac{\epsilon}{\sqrt{d}}\|a\|_2\|b\|_2\|\Sigma^{-1}\|_2,$$

with probability at least $1 - 1/\operatorname{poly}(d)$, as long as $m \gtrsim d(C\log n)^{2k+2}\log^2(n/\epsilon)/\epsilon^2$ for a sufficiently large constant C. Combining with (9), the result holds as long as $m \gtrsim \epsilon^{-2}d\log^c n\max((C\log n)^{2k+2}, d^{\frac{1}{k}})$, for any k. Setting $k = \Theta(\sqrt{\frac{\log d}{\log\log n}})$ gives the result. ◀

Combining Different Matrices. In some cases it can make sense to combine different sketching matrices that satisfy the generalization bound. We defer the details to the full version.

▶ **Theorem 11.** *Let $A \in \mathbb{R}^{n\times d}$, and let $R \in \mathbb{R}^{m\times r}$ and $S \in \mathbb{R}^{r\times n}$ be drawn from distributions of matrices that are ϵ-approximate OSEs and satisfy the generalization bound (6). Then RS satisfies the generalization bound with a constant factor loss in failure probability and approximation factor.*

4 Proof of Lemma 9

Proof. Each column S_i of the subsampled randomized Hadamard transform has the same distribution as σ_iS_i, where σ_i is a random sign. It also has $\langle S_i, S_i\rangle = 1$ for all i and $|\langle S_i, S_j\rangle| \lesssim \frac{\sqrt{\log(1/\delta)}}{\sqrt{r}}$ with probability $1 - \delta$, for any δ and $i \neq j$. See, e.g., [17].

By expanding the following product into a sum, and rearranging terms, we obtain

$$\begin{aligned}
a^\top S^\top S(UU^\top S^\top S)^k b = & \sum_{i_0,j_0,i_1,j_1,\cdots,i_k,j_k} a_{i_0} b_{j_k} \sigma_{i_0}\sigma_{i_1}\cdots\sigma_{i_k}\sigma_{j_0}\sigma_{j_1}\cdots\sigma_{j_k} \\
& \cdot\langle S_{i_0}, S_{j_0}\rangle (UU^\top)_{j_0,i_1}\langle S_{i_1}, S_{j_1}\rangle\cdots(UU^\top)_{j_{k-1},i_k}\langle S_{i_k}, S_{j_k}\rangle \\
= & \sum_{i_0,j_k} a_{i_0} b_{j_k}\sigma_{i_0}\sigma_{j_k} \sum_{j_0,i_1,j_1,\cdots,i_k} \sigma_{i_1}\cdots\sigma_{i_k}\sigma_{j_0}\sigma_{j_1}\cdots\sigma_{j_{k-1}} \\
& \cdot \langle S_{i_0}, S_{j_0}\rangle (UU^\top)_{j_0,i_1}\langle S_{i_1}, S_{j_1}\rangle\cdots(UU^\top)_{j_{k-1},i_k}\langle S_{i_k}, S_{j_k}\rangle \\
= & \sum_{i_0,j_k} \sigma_{i_0}\sigma_{j_k} Z_{i_0,j_k}
\end{aligned}$$

where Z_{i_0,j_k} is defined to be

$$Z_{i_0,j_k} \quad = \quad a_{i_0} b_{j_k} \sum_{\substack{i_1,\cdots i_k \\ j_0,\cdots j_{k-1}}} \prod_{c=1}^{k}\sigma_{i_c}\prod_{c=0}^{k-1}\sigma_{j_c}\cdot\prod_{c=0}^{k}\langle S_{i_c}, S_{j_c}\rangle\prod_{c=1}^{k}(UU^\top)_{i_{c-1},j_c}.$$

Note that Z_{i_0,j_k} is independent of σ_{i_0} and σ_{j_k}. We observe that in the above expression if $i_0 = j_0, i_1 = j_1, \cdots, i_k = j_k$, then the sum over these indices equals $a^\top(UU^\top)\cdots(UU^\top)b = 0$, since $\langle S_{i_c}, S_{j_c}\rangle = 1$ in this case for all c. Moreover, the sum over all indices conditioned on $i_k = j_k$ is equal to 0. Indeed, in this case, the expression can be factored into the form $\zeta \cdot U^\top b$, for some random variable ζ, but $U^\top b = 0$.

Let W be a matrix with $W_{i,j} = \sigma_i\sigma_j Z_{i,j}$. We need Khintchine's inequality:

▶ Fact 12. (Khintchine's Inequality) Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. sign random variables, and let $z_1, \ldots, z_n$ be real numbers. Then there are constants $C, C' > 0$ so that $\Pr[|\sum_{i=1}^n z_i\sigma_i| \geq Ct\|z\|_2] \leq e^{-C't^2}$.

We note that Khintchine's inequality sometimes refers to bounds on the moment of $|\sum_i z_i\sigma_i|$, though the above inequality follows readily by applying a Markov bound to the high moments.

We apply Fact 12 to each column of W, so that if W_i is the i-th column, we have by a union bound that with probability $1 - 1/\text{poly}(n)$, $\|W_i\|_2 = O(\|Z_i\|_2\sqrt{\log n})$ simultaneously for all columns i. It follows that with the same probability, $\|W\|_F^2 = O(\|Z\|_F^2 \log n)$, that is, $\|W\|_F = O(\|Z\|_F\sqrt{\log n})$. We condition on this event in the remainder.

Thus, it remains to bound $\|Z\|_F$. By squaring Z_{i_0,j_0} and using that $\mathbf{E}[\sigma_i\sigma_j] = 1$ if $i = j$ and 0 otherwise, we have,

$$\mathop{\mathbf{E}}_{\sigma}[Z_{i_0,j_k}^2] = a_{i_0}^2 b_{j_k}^2 \sum_{\substack{i_1,\cdots i_k \\ j_0,\cdots j_{k-1}}} \prod_{c=0}^{k}\langle S_{i_c}, S_{j_c}\rangle^2 \prod_{c=1}^{k}(UU^\top)_{i_{c-1},j_c}^2. \tag{10}$$

Due to space considerations, we defer to the full version the proof that

$$\mathop{\mathbf{E}}_{S}[\|Z\|_F^2] \leq (O(d(\log n)/r) + 1)^{k-1}\cdot(d\|b\|_2^2(\log^2 n)/r^2 + \|b\|_2^2(\log n)/r).$$

Note that we also have the bound:

$$(O(d(\log n)/r) + 1)^{k-1} \leq (e^{O(d(\log n)/r)})^{k-1} \leq e^{O(kd(\log n)/r)} \leq O(1),$$

for any $r = \Omega(kd\log n)$.

Having computed the expectation of $\|Z\|_F^2$, we now would like to show concentration. Consider a specific

$$Z_{i_0,j_k} = a_{i_0} b_{j_k}\sum_{i_k}\sigma_{i_k}\langle S_{i_k}, S_{j_k}\rangle\cdots\sum_{j_1}\sigma_{j_1}(UU^\top)_{j_1,i_2}\sum_{i_1}\sigma_{i_1}\langle S_{i_1}, S_{j_1}\rangle\sum_{j_0}\sigma_{j_0}\langle S_{i_0}, S_{j_0}\rangle(UU^\top)_{j_0,i_1}.$$

By Fact 12, for each fixing of i_1, with probability $1 - 1/\text{poly}(n)$, we have

$$\sum_{j_0} \sigma_{j_0} \langle S_{i_0}, S_{j_0} \rangle (UU^\top)_{j_0,i_1} = O(\sqrt{\log n})(\sum_{j_0} \langle S_{i_0}, S_{j_0} \rangle^2 (UU^\top)^2_{j_0,i_1})^{\frac{1}{2}}. \tag{11}$$

Now, we can apply Khintchine's inequality for each fixing of j_1, and combine this with (11). With probability $1 - 1/\text{poly}(n)$, again we have

$$\begin{aligned}
&\sum_{i_1} \sigma_{i_1} \langle S_{i_1}, S_{j_1} \rangle \sum_{j_0} \sigma_{j_0} \langle S_{i_0}, S_{j_0} \rangle (UU^\top)_{j_0,i_1} \\
&= \sum_{i_1} \sigma_{i_1} \langle S_{i_1}, S_{j_1} \rangle O(\sqrt{\log n})(\sum_{j_0} \langle S_{i_0}, S_{j_0} \rangle^2 (UU^\top)^2_{j_0,i_1})^{\frac{1}{2}} \\
&= O(\log n)(\sum_{i_1} \langle S_{i_1}, S_{j_1} \rangle^2 \sum_{j_0} \langle S_{i_0}, S_{j_0} \rangle^2 (UU^\top)^2_{j_0,i_1})^{\frac{1}{2}}.
\end{aligned}$$

Thus, we can apply Khintchine's inequality recursively over all the $2k$ indexes $j_0, i_1, j_1, \cdots, j_{k-1}$, i_k, from which it follows that with probability $1 - 1/\text{poly}(n)$, for each such i_0, j_k, we have $Z^2_{i_0,j_k} = O(\log^k n)\mathop{\mathbf{E}}_S[Z^2_{i_0,j_k}]$, using (10). We thus have with this probability, that $\|Z\|_F^2 = O(\log^k n)\mathop{\mathbf{E}}_S[\|Z\|_F^2]$, completing the proof. ◀

5 Lower bound for ℓ_2 and ℓ_∞ guarantee

We prove a lower bound for the ℓ_2 guarantee, which immediately implies a lower bound for the ℓ_∞ guarantee.

▶ **Definition 13.** Given a matrix $A \in \mathbb{R}^{n\times d}$, vector $b \in \mathbb{R}^n$ and matrix $S \in \mathbb{R}^{r\times n}$, denote $x^* = A^\dagger b$. We say that an algorithm $\mathcal{A}(A, b, S)$ that outputs a vector $x' = (SA)^\dagger Sb$ "succeeds" if the following property holds: $\|x' - x^*\|_2 \lesssim \epsilon \|b\|_2 \cdot \|A^\dagger\|_2 \cdot \|Ax^* - b\|_2$.

▶ **Theorem 14.** *Suppose Π is a distribution over $\mathbb{R}^{m\times n}$ with the property that for any $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, $\Pr_{S\sim\Pi}[\mathcal{A}(A, b, S)$ succeeds $] \geq 19/20$. Then $m \gtrsim \min(n, d/\epsilon^2)$.*

Proof. The proof uses Yao's minimax principle. Let $\mathcal{D}$ be an arbitrary distribution over $\mathbb{R}^{n\times(d+1)}$, then $\mathop{\mathbb{E}}_{(A,b)\sim\mathcal{D}} \mathop{\mathbb{E}}_{S\sim\Pi}[\mathcal{A}(A, b, S)$ succeeds $] \geq 1-\delta$. Switching the order of probabilistic quantifiers, an averaging argument implies the existence of a fixed matrix $S_0 \in \mathbb{R}^{m\times n}$ such that $\mathop{\mathbb{E}}_{(A,b)\sim\mathcal{D}}[\mathcal{A}(A, b, S_0)$ succeeds $] \geq 1 - \delta$. Thus, we must construct a distribution $\mathcal{D}_{\text{hard}}$ such that $\mathop{\mathbb{E}}_{(A,b)\sim\mathcal{D}_{\text{hard}}}[\mathcal{A}(A, b, S_0)$ succeeds $] \geq 1 - \delta$ cannot hold for any $\Pi_0 \in \mathbb{R}^{m\times n}$ which does not satisfy $m = \Omega(d/\epsilon^2)$. The proof can be split into three parts. First, we prove a useful property. Second, we prove a lower bound for the case $\text{rank}(S) \geq d$. Third, we show why $\text{rank}(S) \geq d$ is necessary.

(I) We show that $[SA, Sb]$ are independent Gaussian, if both $[A, b]$ and S are orthonormal matrices. We can rewrite SA in the following sense,

$$\underbrace{S}_{m\times n} \cdot \underbrace{A}_{n\times d} = \underbrace{S}_{m\times n}\underbrace{R}_{n\times n}\underbrace{R^\top}_{n\times n}\underbrace{A}_{n\times d} = S \begin{bmatrix} S^\top & \overline{S}^\top \end{bmatrix} \begin{bmatrix} S \\ \overline{S} \end{bmatrix} A = \begin{bmatrix} I_m & 0 \end{bmatrix} \begin{bmatrix} S \\ \overline{S} \end{bmatrix} A = \begin{bmatrix} I_m & 0 \end{bmatrix} \underbrace{\widetilde{A}}_{n\times d} = \underbrace{\widetilde{A}_m}_{m\times d}$$

where $\overline{S}$ is the complement of the orthonormal basis S, I_m is a $m \times m$ identity matrix, and $\widetilde{A}_m$ is the left $m \times d$ submatrix of $\widetilde{A}$. Thus, using [13] as long as $m = o(\sqrt{n})$ (because of $n = \Omega(d^3)$) the total variation distance between $[SA, Sb]$ and a random Gaussian matrix is small, i.e.,

$$D_{TV}([SA, Sb], H) \leq 0.01 \tag{12}$$

where each entry of H is i.i.d. Gaussian $\mathcal{N}(0, 1/n)$.

(II) Here we prove the theorem in the case when S has rank $r \geq d$ (we will prove this is necessary in part III. Writing $S = U\Sigma V^\top$ in its SVD, we have

$$\underbrace{S}_{m\times n} A = \underbrace{U}_{m\times r}\underbrace{\Sigma}_{r\times r}\underbrace{V^\top}_{r\times n} RR^\top A = U\Sigma G \tag{13}$$

where $R = \begin{bmatrix} V & \overline{V} \end{bmatrix}$. By a similar argument in Equation (12), as long as $r = o(\sqrt{n})$ we have that G also can be approximated by a Gaussian matrix, where each entry is sampled from i.i.d. $\mathcal{N}(0, 1/n)$. Similarly, $Sb = U\Sigma h$, where h also can be approximated by a Gaussian matrix, where each entry is sampled from i.i.d. $\mathcal{N}(0, 1/n)$.

Since U has linearly independent columns, $(U\Sigma G)^\dagger U\Sigma h = (\Sigma G)^\dagger U^\top U\Sigma h = (\Sigma G)^\dagger \Sigma h$.

The $r \times d$ matrix G has SVD $G = \underbrace{R}_{r\times d}\underbrace{\widetilde{\Sigma}}_{d\times d}\underbrace{T}_{d\times d}$, and applying the pseudo-inverse property again, we have

$$\begin{aligned}\|(SA)^\dagger Sb\|_2 &= \|(\Sigma G)^\dagger \Sigma h\|_2 = \|(\Sigma R\widetilde{\Sigma}T)^\dagger \Sigma h\|_2 = \|T^\dagger(\Sigma R\widetilde{\Sigma})^\dagger \Sigma h\|_2 = \|(\Sigma R\widetilde{\Sigma})^\dagger \Sigma h\|_2 \\ &= \|\widetilde{\Sigma}^\dagger(\Sigma R)^\dagger \Sigma h\|_2,\end{aligned}$$

where the the first equality follows by Equation (13), the second equality follows by the SVD of G, the third and fifth equality follow by properties of the pseudo-inversewhen T has orthonormal rows and $\widetilde{\Sigma}$ is a diagonal matrix, and the fourth equality follows since $\|T^\dagger\|_2 = 1$ and T is an orthonormal basis.

Because each entry of $G = R\widetilde{\Sigma}T \in \mathbb{R}^{r\times d}$ is sampled from an i.i.d. Gaussian $\mathcal{N}(0, 1)$, using the result of [24] we can give an upper bound for the maximum singular value of G: $\|\widetilde{\Sigma}\| \lesssim \sqrt{\frac{r}{n}}$ with probability at least .99. Thus,

$$\|\widetilde{\Sigma}^\dagger(\Sigma R)^\dagger \Sigma h\|_2 \geq \sigma_{\min}(\widetilde{\Sigma}^\dagger) \cdot \|(\Sigma R)^\dagger \Sigma h\|_2 = \sigma_{\max}^{-1}(\widetilde{\Sigma})\|(\Sigma R)^\dagger \Sigma h\|_2 \gtrsim \sqrt{n/r}\|(\Sigma R)^\dagger \Sigma h\|_2.$$

Because h is a random Gaussian vector which is independent of $(\Sigma R)^\dagger\Sigma$, $\mathbf{E}_h[\|(\Sigma R)^\dagger \Sigma h\|_2^2] = \frac{1}{n} \cdot \|(\Sigma R)^\dagger \Sigma\|_F^2$, where each entry of h is sampled from i.i.d. Gaussian $\mathcal{N}(0, 1/n)$. Then, using the Pythagorean Theorem, $\|(\Sigma R)^\dagger\Sigma\|_F^2 = \|(\Sigma R)^\dagger \Sigma RR^\top\|_F^2 + \|(\Sigma R)^\dagger\Sigma(I - RR^\top)\|_F^2 \geq \|(\Sigma R)^\dagger \Sigma RR^\top\|_F^2 = \|(\Sigma R)^\dagger \Sigma R\|_F^2 = \text{rank}(\Sigma R) = \text{rank}(SA) = d$. Thus, $\|x' - x^*\|_2 \gtrsim \sqrt{d/r} \geq \sqrt{d/m} = \epsilon$.

(III) Now we show that we can assume that $\text{rank}(S) \geq d$.

We sample A, b based on the following distribution $\mathcal{D}_{\text{hard}}$: with probability $1/2$, A, b are sampled from $\mathcal{D}_1$; with probability $1/2$, A, b are sampled from $\mathcal{D}_2$. In distribution $\mathcal{D}_1$, A is a random orthonormal basis and d is always orthogonal to A. In distribution $\mathcal{D}_2$, A is a $d \times d$ identity matrix in the top-d rows and 0s elsewhere, while b is a random unit vector. Then, for any (A, b) sampled from $\mathcal{D}_1$, S needs to work with probability at least $9/10$. Also for any (A, b) sampled from $\mathcal{D}_2$, S needs to work with probability at least $9/10$. The latter two statements follow since overall S succeeds on $\mathcal{D}_{\text{hard}}$ with probability at least $19/20$.

Consider the case where A, b are sampled from distribution $\mathcal{D}_2$. Then $x^* = b$ and $\text{OPT} = 0$. Then consider x' which is the optimal solution to $\min_x \|SAx - Sb\|_2^2$, so $x' = (SA)^\dagger Sb = (S_L)^\dagger S_L b$, where S can be decomposed into two matrices $S_L \in \mathbb{R}^{r\times d}$ and $S_R \in \mathbb{R}^{r\times(n-d)}$, $S = \begin{bmatrix} S_L & S_R \end{bmatrix}$. Plugging x' into the original regression problem, $\|Ax' - b\|_2^2 = \|A(S_L)^\dagger S_L b - b\|_2^2$, which is at most $(1 + \epsilon)\,\text{OPT} = 0$. Thus $\text{rank}(S_L)$ is d. Since S_L is a submatrix of S, the rank of S is also d. ◀

References

1 Patrick Amestoy, Iain S. Duff, Jean-Yves L'Excellent, Yves Robert, François-Henry Rouet, and Bora Uçar. On computing inverse entries of a sparse matrix in an out-of-core environment. *SIAM J. Scientific Computing*, 34(4), 2012.

2 Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM J. Scientific Computing*, 32(3):1217–1236, 2010.

3 Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *STOC*, 2015.

4 Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

5 Kenneth L Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

6 Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *SODA*, pages 278–287, 2016.

7 Michael B Cohen, Cameron Musco, and Christopher Musco. Ridge leverage scores for low-rank approximation. *SODA*, 2017.

8 Michael B Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In *ICALP*, 2016.

9 Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.*, 9(3):251–280, 1990.

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

11 Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

12 François Le Gall. Powers of tensors and fast matrix multiplication. In *International Symposium on Symbolic and Algebraic Computation, ISSAC '14, Kobe, Japan, July 23-25, 2014*, pages 296–303, 2014.

13 Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):1497–1529, 2006.

14 Jeff Leek. Prediction: the lasso vs. just using the top 10 predictors. In *simplystats, February 23.* `http://simplystatistics.org/2012/02/23/prediction-the-lasso-vs-just-using-the-top-10/`, 2012.

15 Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 127–136, 2013.

16 Song Li, Shaikh S. Ahmed, Gerhard Klimeck, and Eric Darve. Computing entries of the inverse of a sparse matrix using the FIND algorithm. *J. Comput. Physics*, 227(22):9408–9427, 2008.

17 Yichao Lu, Paramveer Dhillon, Dean Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *NIPS*, 2013.

18 Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100. ACM, 2013.

19 Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 117–126. IEEE, 2013.

20 Jelani Nelson and Huy L. Nguyên. Lower bounds for oblivious subspace embeddings. In *ICALP*, pages 883–894, 2014.

21 Dimitris Papailiopoulos, Anastasios Kyrillidis, and Christos Boutsidis. Provable deterministic leverage score sampling. In *KDD*, pages 997–1006. ACM, 2014.

22 Ryan Rifkin, Gene Yeo, and Tomaso Poggio. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

23 Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

24 Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

25 Virginia Vassilevska Williams. Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 887–898, 2012.

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

27 Peng Zhang and Jing Peng. Svm vs regularized least squares classification. In *ICPR (1)*, pages 176–179, 2004.

28 Anastasios Zouzias and Nikolaos M. Freris. Randomized extended kaczmarz for solving least squares. *SIAM J. Matrix Analysis Applications*, 34(2):773–793, 2013.