Sketching Algorithms and Lower Bounds for Ridge Regression

Praneeth Kacham [1] **David P. Woodruff** [1]

Abstract

We give a sketching-based iterative algorithm that computes a $1+\varepsilon$ approximate solution for the ridge regression problem $\min_x \|Ax-b\|_2^2 + \lambda\|x\|_2^2$ where $A \in \mathbb{R}^{n\times d}$ with $d \geq n$. Our algorithm, for a constant number of iterations (requiring a constant number of passes over the input), improves upon earlier work (Chowdhury et al., 2018) by requiring that the sketching matrix only has a weaker Approximate Matrix Multiplication (AMM) guarantee that depends on ε, along with a constant subspace embedding guarantee. The earlier work instead requires that the sketching matrix has a subspace embedding guarantee that depends on ε. For example, to produce a $1+\varepsilon$ approximate solution in 1 iteration, which requires 2 passes over the input, our algorithm requires the OSNAP embedding to have $m = O(n\sigma^2/\lambda\varepsilon)$ rows with a sparsity parameter $s = O(\log(n))$, whereas the earlier algorithm of Chowdhury et al. (2018) with the same number of rows of OSNAP requires a sparsity $s = O(\sqrt{\sigma^2/\lambda\varepsilon}\cdot\log(n))$, where $\sigma = \|A\|_2$ is the spectral norm of the matrix A. We also show that this algorithm can be used to give faster algorithms for kernel ridge regression. Finally, we show that the sketch size required for our algorithm is essentially optimal for a natural framework of algorithms for ridge regression by proving lower bounds on oblivious sketching matrices for AMM. The sketch size lower bounds for AMM may be of independent interest.

[1]Computer Science Department, Carnegie Mellon University. Correspondence to: Praneeth Kacham <pkacham@cs.cmu.edu>, David P. Woodruff <dwoodruf@andrew.cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Given a matrix $A \in \mathbb{R}^{n\times d}$, a vector $b \in \mathbb{R}^n$, and a parameter $\lambda \geq 0$, the ridge regression problem is defined as:

$$\min_x \|Ax-b\|_2^2 + \lambda\|x\|_2^2.$$

Throughout the paper, we assume $n \leq d$, and that x^* is the optimal solution for the problem. Let Opt be the optimal value for the above problem. Earlier work (Chowdhury et al., 2018) gives an iterative algorithm using so-called subspace embeddings. The following theorem states their results when their algorithm is run for 1 iteration. Note that their algorithm is more general and when run for t iterations, the error is proportional to ε^t.

Theorem 1.1 (Theorem 1 of Chowdhury et al. (2018))**.** *Given $A \in \mathbb{R}^{n\times d}$, let $V \in \mathbb{R}^{d\times n}$ be an orthonormal basis for the rowspace of matrix A. If $S \in \mathbb{R}^{m\times d}$ is a matrix which satisfies*

$$\|V^\mathsf{T}S^\mathsf{T}SV - I_n\|_2 \leq \varepsilon/2, \tag{1}$$

then $\tilde{x} = A^\mathsf{T}(AS^\mathsf{T}SA^\mathsf{T} + \lambda I_n)^{-1}b$ satisfies

$$\|\tilde{x} - x^*\|_2 \leq \varepsilon\|x^*\|_2.$$

A matrix S which satisfies (1) is called an $\varepsilon/2$ subspace embedding for the column space of V, since for any y in colspan(V), we have $(1-\varepsilon/2)\|y\|_2^2 \leq \|Sy\|_2^2 \leq (1+\varepsilon/2)\|y\|_2^2$. We also frequently drop the term "column space" and say S is an $\varepsilon/2$ subspace embedding for the matrix V itself.

There are many *oblivious* and *non-oblivious* constructions of subspace embeddings. As the name suggests, *oblivious* subspace embedding (OSE) constructions do not depend on the matrix V that is to be embedded. OSEs specify a distribution $\mathcal{S}$ such that for any arbitrary matrix V, a random matrix $\mathbf{S}$ drawn from the distribution $\mathcal{S}$ is an ε subspace embedding for V with probability $\geq 1-\delta$. On the other hand, non-oblivious constructions compute a distribution $\mathcal{S}$ that depends on the matrix V that is to be embedded. See the survey by Woodruff (2014) for an overview.

In many cases, such as in streaming, it is important that the sketch used is *oblivious*, since matrix-dependent subspace embedding constructions may need to read the entire input

matrix first. Oblivious sketches also allow turnstile updates to the matrix A in a stream. In the turnstile model of streaming, we receive updates of the form $((i,j),v)$ which update $A_{i,j}$ to $A_{i,j}+v$. In our paper we focus on algorithms for ridge regression that use *oblivious* sketching matrices.

To satisfy (1), using CountSketch (Clarkson & Woodruff, 2017; Meng & Mahoney, 2013), we can obtain a sketching dimension of $m = O(n^2/\varepsilon^2)$ for which the matrix SA^T can be computed in $O(\mathtt{nnz}(A))$ time, where $\mathtt{nnz}(A)$ denotes the number of nonzero entries in the matrix A. Using OSNAP embeddings (Nelson & Nguyen, 2013; Cohen, 2016), we can obtain a sketching dimension of $m = O(n^{1+\gamma}\log(n)/\varepsilon^2)$ for which the matrix SA^T can be computed in time $O(\mathtt{nnz}(A)/\gamma\varepsilon)$. For $\gamma = O(1/\log(n))$, we have $m = O(n\log(n)/\varepsilon^2)$ with SA^T that can be computed in time $O(\mathtt{nnz}(A)\log(n)/\varepsilon)$. We can see that there is a tradeoff between CountSketch and OSNAP — one has a smaller sketching dimension while the other is faster to apply to a given matrix. If $t_{SA^\intercal}$ is the time required to compute SA^T, then $\tilde{x}$ in Theorem 1.1 can be computed in time $O(\mathtt{nnz}(A) + t_{SA^\intercal} + mn^{\omega-1} + n^\omega)$ where ω is the matrix multiplication constant. Thus it is important to have both a small $t_{SA^\intercal}$ and small m to obtain fast running times.

When allowed $O(\log(1/\varepsilon))$ passes over the input matrix A, the algorithm of Chowdhury et al. (2018) produces an ε relative error solution using only a constant, say $1/2$ subspace embedding. When only $O(1)$ passes are allowed over the input, their algorithm requires a $\delta = f(\varepsilon)$ subspace embedding to obtain ε error solutions. As seen above this leads to either a high value of m or a high value of $t_{SA^\intercal}$.

We show that we only need a simpler Approximate Matrix Multiplication (AMM) guarantee, along with a constant subspace embedding, instead of requiring S to be an $\varepsilon/2$ subspace embedding.

Definition 1.2 (AMM)**.** Given matrices A and B of appropriate dimensions, a matrix S satisfies the ε-AMM property for (A,B) if

$$\|A^\mathsf{T}S^\mathsf{T}SB - A^\mathsf{T}B\|_\mathsf{F} \le \varepsilon\|A\|_\mathsf{F}\|B\|_\mathsf{F}.$$

We now state the guarantees of our algorithm (Algorithm 1) for 1 iteration, requiring 2 passes over the matrix A.

Theorem 1.3. *If $\mathbf{S}$ is a random matrix such that for any fixed $d\times n$ orthonormal matrix V and a vector r, with probability $\ge 9/10$,*

$$\|V^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}V - I_n\|_2 \le 1/2$$

and

$$\|V^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}Vr - r\|_2 \le (\varepsilon/2\sqrt{n})\|V\|_\mathsf{F}\|Vr\|_2 = (\varepsilon/2)\|r\|_2, \tag{2}$$

then $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I_n)^{-1}b$ satisfies $\|\tilde{x}-x^*\|_2 \le \varepsilon\|x^*\|_2$ with probability $\ge 9/10$.

We show that the OSNAP distribution satisfies both of these two properties with a sketching dimension of $r = O(n\log(n) + n/\varepsilon^2)$ and with SA^T that can be computed in $O(\mathtt{nnz}(A)\cdot\log(n))$ time. Note that our algorithm (Algorithm 1) is also more general, and when run for t iterations, the error is proportional to ε^t. Our algorithm differs from that of Chowdhury et al. (2018) in that our algorithm needs a fresh sketching matrix in each iteration whereas their algorithm only needs one sketching matrix across iterations.

Many natural problems in the streaming literature have been studied specifically with 2 passes (Chen et al., 2021; Konrad & Naidu, 2021; Assadi & Raz, 2020; Brody & Woodruff, 2011). Also in the case of federated learning, where minimizing the number of rounds of communication is important (Park et al., 2021), the smaller sketch sizes required by our algorithm (Algorithm 1) gives an improvement over the algorithm of Chowdhury et al. (2018).

We can also bound the cost of $\tilde{x}$ computed by our algorithm. For any $x\in\mathbb{R}^d$, let $\mathrm{cost}(x) = \|Ax-b\|_2^2 + \lambda\|x\|_2^2$. Bounds on $\|\tilde{x}-x^*\|_2$ also let us obtain an upper bound on $\mathrm{cost}(\tilde{x})$. It can be shown that for any vector x, $\mathrm{cost}(\tilde{x}) = \mathrm{Opt} + \|A(x^*-\tilde{x})\|_2^2 + \lambda\|x^*-\tilde{x}\|_2^2$. Thus, $\|\tilde{x}-x^*\|_2 \le \varepsilon\|x^*\|_2$ implies that $\mathrm{cost}(\tilde{x}) = \mathrm{Opt} + (\sigma^2+\lambda)\varepsilon^2\|x^*\|_2^2 \le (1 + (1+\sigma^2/\lambda)\varepsilon^2)\mathrm{Opt}$. Throughout the paper, we are most interested in the case $\sigma^2 \ge \lambda$, as it is when $\mathrm{cost}(\tilde{x})$ could be much higher than Opt. Setting $\varepsilon = O(\sqrt{\delta\lambda/\sigma^2})$, we obtain that the solution $\tilde{x}$ returned by Theorem 1.3 is a $1+\delta$ approximation.

We also show that our algorithm can be used to obtain approximate solutions to Kernel Ridge Regression with a polynomial kernel. We show that instantiating the construction of Ahle et al. (2020) with appropriate sketching matrices gives a fast way to apply sketches, satisfying the subspace embedding and AMM properties, to the matrix $\phi(A)$, where the i-th row of the matrix $\phi(A)$ is given by $A_{i*}^{\otimes p}$.

1.1. Lower bounds for Ridge Regression

It can be seen that the optimal solution $x^* = A^\mathsf{T}(AA^\mathsf{T} + \lambda I_n)^{-1}b$. Our algorithm, for one iteration, is simply to compute $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I_n)^{-1}b$ for a matrix $\mathbf{S}$ that satisfies the requirements in Theorem 1.3. All the algorithm does is substitute the expensive matrix product AA^T, which can take $O(n\cdot\mathtt{nnz}(A))$ time to compute, with the matrix product $A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T}$, which only takes $t_{\mathbf{S}A^\intercal} + mn^{\omega-1}$ time to compute. Thus, constructing "good" distributions for which $\tilde{x}$ is a $1+\varepsilon$ approximation seems to be the most natural way to obtain fast algorithms for ridge regression. As discussed previously, OSNAP matrices with $m = O(n\log(n) + n\sigma^2/\lambda\varepsilon)$ and having near-optimal $t_{\mathbf{S}A^\intercal} = \tilde{O}(\mathtt{nnz}(A))$ can be used to compute a solution $\tilde{x}$ that is a $1+\varepsilon$ approximation. We show that for a large class of nice-enough distributions over $m\times d$ matrices $\mathcal{S}$, if

$\mathbf{S} \sim \mathcal{S}$ satisfies that $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I)^{-1}b$ is a $1+\varepsilon$ approximation with high probability, then $r = \Omega(n\sigma^2/\lambda\varepsilon)$. This shows that OSNAP matrices have both a near-optimal sketching dimension r and near-optimal time $t_{\mathbf{S}A^\mathsf{T}}$. We show the lower bound by showing that for any "nice" distribution $\mathcal{S}$ for which $\tilde{x}$ is a $1+\varepsilon$ approximation with high probability, the distribution must also satisfy an Approximate Matrix Multiplication (AMM) guarantee, i.e., for any matrix B, for $\mathbf{S} \sim \mathcal{S}$, $\|B^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}B - B^\mathsf{T}B\|_\mathsf{F}$ must be small with high probability. We then show a lower bound on m for any distribution $\mathcal{S}$ which satisfies the AMM guarantee. Here we demonstrate our techniques in the simple case of $n = 1$. Without loss of generality, we assume $\lambda = 1$.

Consider the ridge regression problem $\min_x (a^\mathsf{T}x - b)^2 + \|x\|_2^2$, where a is an arbitrary d-dimensional vector. We have $\tilde{x} = a(a^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}a + 1)^{-1}b$ and

$$\text{cost}(\tilde{x}) = \left(\frac{\|a\|_2^2 b}{\|\mathbf{S}a\|_2^2 + 1} - b\right)^2 + \frac{\|a\|_2^2}{(\|\mathbf{S}a\|_2^2 + 1)^2}b^2$$

whereas $\text{Opt} = b^2/(\|a\|_2^2 + 1)$. For $\|a\|_2 \geq 100/\sqrt{\varepsilon}$, it turns out that unless $(1 - \sqrt{\varepsilon}/\|a\|_2)\|a\|_2^2 \leq \|\mathbf{S}a\|_2^2 \leq (1 + \sqrt{\varepsilon}/\|a\|_2)\|a\|_2^2$, we will have $\text{cost}(\tilde{x}) \geq (1+\varepsilon/2)\text{Opt}$. Thus for $\tilde{x}$ to be a $1+\varepsilon/2$ approximation with probability $\geq 99/100$ for any arbitrary a, it must be the case that with probability $\geq 99/100$, $|a^\mathsf{T}a - a^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}a| = |\|a\|_2^2 - \|\mathbf{S}a\|_2^2| \leq (\sqrt{\varepsilon}/\|a\|_2)\|a\|_2^2$ i.e., $\mathbf{S}$ must satisfy the AMM property with parameter $\sqrt{\varepsilon}/\|a\|_2$. We show an $\Omega(1/\delta^2)$ lower bound for any distribution which satisfies the δ-AMM property, which gives a lower bound of $\Omega(\|a\|_2^2/\varepsilon)$ for ridge regression for $n = 1$.

For the case of general n, we show that any "nice" distribution $\mathcal{S}$ that gives $1+\varepsilon$ approximate solutions for ridge regression must satisfy the $\sqrt{\varepsilon/n\sigma^2}$-AMM guarantee, which by using the lower bound for AMM, gives an $\Omega(n\sigma^2/\varepsilon)$ lower bound for ridge regression.

To prove the lower bound, we crucially use the fact that the sketching distribution $\mathcal{S}$ must satisfy that $\tilde{x}$ is a $1+\varepsilon$ approximation for *any* particular ridge regression problem instance (A, b) with high probability.

1.2. Lower bounds for AMM

We prove the following lower bound for oblivious sketching matrices that give AMM guarantees.

Theorem 1.4 (Informal)**.** *If $\mathcal{S}$ is a distribution over $m \times d$ matrices such that for any $n \times d$ matrix A, $\mathbf{S} \sim \mathcal{S}$ satisfies with probability $\geq 99/100$, that*

$$\|A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} - AA^\mathsf{T}\|_\mathsf{F} \leq \delta\|A\|_\mathsf{F}\|A^\mathsf{T}\|_\mathsf{F},$$

for $\delta \leq c/\sqrt{n}$, then $m = \Omega(1/\delta^2)$ where $c > 0$ is a small enough universal constant.

To the best of our knowledge, this is the first tight lower bound on the dimension of oblivious sketching matrices for AMM. The lower bound is tight up to constant factors as the CountSketch distribution with $m = O(1/\delta^2)$ rows has the above property. Note that for $\delta = \varepsilon/n$, the distribution $\mathcal{S}$ as in the above theorem satisfies that for any $d \times n$ orthonormal matrix V, with probability $\geq 99/100$,

$$\|V^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}V - I_n\|_\mathsf{F} \leq (\varepsilon/n)\|V\|_\mathsf{F}^2 = \varepsilon.$$

Thus, a distribution $\mathcal{S}$ that has the ε/n-AMM property also has the ε-subspace embedding property. Nelson & Nguyen (2014) gives an $\Omega(n/\varepsilon^2)$ lower bound for such distributions, thus giving an $\Omega(1/(\delta^2 n))$ lower bound for δ-AMM for small enough δ. The above theorem gives a stronger $\Omega(1/\delta^2)$ lower bound.

We now give a brief overview of our proof for $n = 1$. Consider $a \in \mathbb{R}^d$ to be a fixed unit vector and let $\mathcal{S}$ be a distribution supported on $r \times d$ matrices as in the above theorem. Then we have $\Pr_{\mathbf{S}\sim\mathcal{S}}[|a^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}a - 1| \leq \delta] \geq 0.99$. Let $U\Sigma V^\mathsf{T}$ be the singular value decomposition of $\mathbf{S}$ with $\Sigma \in \mathbb{R}^{r\times r}$. Without loss of generality, we can assume that V^T is independent of Σ and that V^T is a uniformly random orthonormal matrix. This follows from the fact that if $\mathbf{S}$ is an *oblivious* AMM sketch, then $\mathbf{SQ}$ is also an *oblivious* AMM sketch, where $\mathbf{Q}$ is a uniformly random $d \times d$ orthogonal matrix independent of $\mathbf{S}$. Thus, we have $\Pr_{\Sigma,V^\mathsf{T}}[|a^\mathsf{T}V\Sigma^2V^\mathsf{T}a - 1| \leq \delta] \geq 0.99$, where Σ and V^T are random matrices that correspond to the AMM sketch $\mathbf{S}$, as described.

Jiang & Ma (2017) show that if $m = o(d)$, then the total variation distance between $V^\mathsf{T}a$ and $(1/\sqrt{d})\mathbf{g}$ is small, where $\mathbf{g}$ is an m dimensional vector with independent Gaussian entries. Thus, we obtain that $\Pr_{\Sigma,\mathbf{g}}[|(1/d)\mathbf{g}^\mathsf{T}\Sigma^2\mathbf{g} - 1| \leq \delta] = \Pr_{\Sigma,\mathbf{g}}[|(1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_i^2 - 1| \leq \delta] \geq 0.95$.

If $\sum_{i=1}^m \sigma_i^2 \leq d/200$, then $(1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_i^2 \leq 1/2$ with probability ≥ 0.99 by Markov's inequality. So, $\Pr_\Sigma[\sum_{i=1}^m \sigma_i^2 \leq d/200]$ must be small. On the other hand, $\text{Var}((1/d)\sum_i \sigma_i^2\mathbf{g}_i^2) = (2/d^2)\sum_{i=1}^m \sigma_i^4$. Thus for $(1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_i^2$ to concentrate in the interval $(1-\delta, 1+\delta)$, we would expect $\sqrt{\text{Var}((1/d)\sum_i \sigma_i^2\mathbf{g}_i^2)} \approx \delta$, which implies $(2/d^2)\sum_{i=1}^m \sigma_i^4 \approx \delta^2$. Thus, with a reasonable probability, it must be simultaneously true that $d/200 \leq \sum_{i=1}^m \sigma_i^2$ and $\sum_{i=1}^m \sigma_i^4 \approx d^2\delta^2/2$. Then,

$$d^2/(200)^2 \leq \left(\sum_{i=1}^m \sigma_i^2\right) \leq m\sum_{i=1}^m \sigma_i^4 \approx md^2\delta^2/2,$$

thus obtaining $m \gtrsim \Omega(1/\delta^2)$. We extend this proof idea to the general case of $n \geq 1$.

Non-asymptotic upper bounds on the total variation (TV) distance between Gaussian matrices and sub-matrices of random orthogonal matrices obtained in recent works (Jiang

& Ma, 2017; Li & Woodruff, 2021) let us replace the matrices that are harder to analyze with Gaussian matrices in our proof of the lower bound for AMM. We believe this technique could be helpful in proving tight lower bounds for other types of sketching guarantees.

1.3. Other Contributions

We also show lower bounds on the communication complexity for approximating the *optimal value* of the ridge regression problem, which is a different, but related problem, to the computation of $1+\varepsilon$ approximate *solutions* which is studied in this paper. We obtain an $\Omega(1/\varepsilon^2)$ bit lower bound for ridge regression when $\sigma^2/\lambda \approx 1$. The hard instance is a two-party communication game on a $2 \times d$ matrix, for which one party has the first row and the other party has the second row. Surprisingly, if each party has $d/2$ columns of the design matrix $A = [A_1\ A_2]$, then they can compute the exact optimal value if the first party communicates $A_1 A_1^\mathsf{T}$ to the second party using $O(n^2)$ words of communication. This suggests that the turnstile streaming setting is harder than the column arrival setting for streaming algorithms for ridge regression. We stress that our algorithm to compute a $(1+\varepsilon)$-approximate solution to ridge regression works in the turnstile streaming setting by maintaining $\mathbf{S}A^\mathsf{T}$ in a stream. Nevertheless, to output the d dimensional solution $\tilde{x}$, our algorithm needs to compute *one* matrix-vector product with A^T at the end of the stream, necessitating a second pass over the stream.

In contrast, we obtain $\Omega(d)$ bit communication complexity lower bounds for the Lasso and square-root Lasso objectives, even for computing $1+c$ approximations to the optimum value for a small enough constant c. Fast algorithms for these objectives seem to be harder to find and as the lower bounds indicate, there may not be sketching-based algorithms for these problems.

We defer most of the proofs to the supplementary material. We also include an experiment there comparing the time required by our algorithm to the time required to compute the exact solution for ridge regression.

2. Preliminaries

For $n \in \mathbb{Z}$, $[n]$ denotes the set $\{1, \ldots, n\}$. Given a matrix $A \in \mathbb{R}^{n\times d}$, let $\mathtt{nnz}(A)$ denote the number of nonzero entries in A. For the matrix A, $\|A\|_\mathsf{F}$ denotes the Frobenius norm $(\sum_{i,j} A_{i,j}^2)^{1/2}$ and $\|A\|_2$ denotes the spectral (operator) norm $\max_{x\neq 0} \|Ax\|_2/\|x\|_2$. When there is no ambiguity, throughout the paper we use σ to denote $\|A\|_2$. Let $A = U\Sigma V^\mathsf{T}$ be the Singular Value Decomposition (SVD) with $U \in \mathbb{R}^{n\times\rho}$, $\Sigma \in \mathbb{R}^{\rho\times\rho}$, and $V \in \mathbb{R}^{d\times\rho}$, where $\rho = \mathrm{rank}(A)$. For arbitrary matrices M, N, the symbol t_{MN} denotes the time required to compute the product MN.

We use uppercase symbols $A, U, V, S, \ldots$ to denote matrices and lowercase symbols $a, b, u, v, \ldots$ to denote vectors. For a matrix A, A_{i*} (A_{*i}) denotes the i-th row (column). We use boldface symbols $\mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{r}, \ldots$ to stress that these objects are random and are explicitly sampled from an appropriate distribution.

Definition 2.1 (Approximate Matrix Multiplication)**.** Given an integer d, we say that an $m \times d$ random matrix $\mathbf{S}$ has the (ε, δ)-AMM property if for any matrices A and B with d rows, we have that

$$\|A^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}B - A^\mathsf{T}B\|_\mathsf{F} \leq \varepsilon\|A\|_\mathsf{F}\|B\|_\mathsf{F}.$$

with probability $\geq 1-\delta$ over the randomness of $\mathbf{S}$.

We usually drop δ from the notation by picking it to be a small enough constant.

Definition 2.2 (Oblivious Subspace Embeddings)**.** Given an integer d, an $m \times d$ random matrix $\mathbf{S}$ is an (ε, δ)-OSE for n-dimensional subspaces if for any arbitrary $d \times n$ matrix A, with probability $\geq 1-\delta$, simultaneously for all vectors x,

$$\|\mathbf{S}Ax\|_2^2 \in (1 \pm \varepsilon)\|Ax\|_2^2.$$

For both OSEs and distributions satisfying the (ε, δ)-AMM property, two major parameters of importance are the size of the sketch (m), and the time to compute $\mathbf{S}A$ ($t_{\mathbf{S}A}$). See Woodruff (2014) and the references therein for several OSE constructions and their corresponding parameters.

3. Iterative Algorithm for Ridge Regression

The following theorem describes the guarantees of the solution $\hat{x}$ returned by Algorithm 1.

Theorem 3.1. *If Algorithm 1 samples independent sketching matrices $\mathbf{S}_j \in \mathbb{R}^{m\times d}$ for all $j \in [t]$ satisfying the properties*

1. *with probability $\geq 1 - 1/(20t)$, for all vectors x, $\|\mathbf{S}_j A^\mathsf{T} x\|_2^2 \in (1 \pm 1/2)\|A^\mathsf{T}x\|_2^2$, and*
2. *for all arbitrary matrices M, N, with probability $\geq 1 - 1/(20t)$,*

$$\|M^\mathsf{T}\mathbf{S}_j^\mathsf{T}\mathbf{S}_j N - M^\mathsf{T}N\|_\mathsf{F} \leq \sqrt{\varepsilon/4n}\|M\|_\mathsf{F}\|N\|_\mathsf{F},$$

then with probability $\geq 9/10$, $\|\hat{x} - x^*\|_2 \leq (\sqrt{\varepsilon})^t\|x^*\|_2$ and further cost$(\hat{x}) \leq (1 + (\sigma^2/\lambda + 1)\varepsilon^t)$Opt.

We prove a few lemmas which give intuition about the algorithm before proving the above theorem.

After $i-1$ iterations of the algorithm, $\sum_{j=1}^{i-1} \tilde{x}^{(j)}$ is the estimate for the optimum solution x^*. At a high level, in the i-th iteration, the algorithm is trying to compute an

Algorithm 1 RIDGEREGRESSION

Input: $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^d, t \in \mathbb{Z}, \varepsilon, \lambda > 0$
Output: $\hat{x} \in \mathbb{R}^d$
$b^{(0)} \leftarrow b$, $\tilde{x}^{(0)} \leftarrow 0_d$, $y^{(0)} \leftarrow 0_n$
for $j = 1, \ldots, t$ **do**
  $b^{(j)} \leftarrow b^{(j-1)} - \lambda y^{(j-1)} - A\tilde{x}^{(j-1)}$
  $\mathbf{S}_j \leftarrow 1/2$ subspace embedding for the rowspace of A and has the $\sqrt{\varepsilon/4n}$ AMM property
  $y^{(j)} \leftarrow (A\mathbf{S}_j^\mathsf{T}\mathbf{S}_j A^\mathsf{T} + \lambda I)^{-1} b^{(j)}$
  $\tilde{x}^{(j)} \leftarrow A^\mathsf{T} y^{(j)}$
end for
$\hat{x} \leftarrow \sum_{j=1}^t \tilde{x}^{(j)}$
return $\hat{x}$

approximation to the difference $x^* - \sum_{j=1}^{i-1} \tilde{x}_j$ by computing an approximate solution to the problem

$$\min_x \|A(x + \sum_{j=1}^{i-1} \tilde{x}_j) - b\|_2^2 + \lambda\|x + \sum_{j=1}^{i-1} \tilde{x}_j\|_2^2.$$

Let $x^{*(j)} = A^\mathsf{T}(AA^\mathsf{T} + \lambda I)^{-1} b^{(j)}$. The following lemma shows that the solution to the above problem is $x^{*(i)}$.

Lemma 3.2. *For all i, $x^* = x^{*(i)} + \sum_{j=1}^{i-1} \tilde{x}^{(j)}$.*

Proof. Let $f(x) = \|A(x + \sum_{j=1}^{i-1} \tilde{x}^{(j)}) - b\|_2^2 + \lambda\|x + \sum_{j=1}^{i-1} \tilde{x}^{(j)}\|_2^2$ and z be the solution realizing $\min_x f(x)$. We have $\nabla_x f(x)\,|_{x=z} = 0$ giving $z = (A^\mathsf{T}A + \lambda I)^{-1}(A^\mathsf{T}b - (A^\mathsf{T}A + \lambda I)\sum_{j=1}^{i-1} \tilde{x}^{(j)})$.

Noting that $\tilde{x}^{(j)} = A^\mathsf{T} y^{(j)}$ for all j and that for all i, $b^{(i)} = b - \lambda \sum_{j=1}^{i-1} y^{(j)} - \sum_{j=1}^{i-1} A\tilde{x}^{(j-1)}$, we obtain that $z = (A^\mathsf{T}A + \lambda I)^{-1} A^\mathsf{T} b^{(i)}$. Now using the matrix identity $(A^\mathsf{T}A + \lambda I)^{-1}A^\mathsf{T} = A^\mathsf{T}(AA^\mathsf{T} + \lambda I)^{-1}$, we get $z = x^{*(i)}$ is the optimal solution to $\min_x f(x)$.

As $x^{*(i)}$ is the optimal solution, it is also clear that $x^* = x^{*(i)} + \sum_{j=1}^{i-1} \tilde{x}^{(j)}$ since otherwise x^* is not the optimal solution for the original ridge regression problem, which is a contradiction. □

So by the end of the $(j-1)$-th iteration, the estimate to x^* is off by $x^{*(j)}$. The algorithm is approximating $x^{*(j)} = A^\mathsf{T}(AA^\mathsf{T} + \lambda I)^{-1} b^{*(j)}$ with $\tilde{x}^{(j)} = A^\mathsf{T}(A\mathbf{S}_j^\mathsf{T}\mathbf{S}_j A^\mathsf{T} + \lambda I)^{-1} b^{*(j)}$. The following lemma gives the error of this approximation assuming that the sketching matrix $\mathbf{S}_j$ has both the subspace embedding and AMM properties. This is the part where our proof differs from that of the proof of Chowdhury et al. (2018).

Lemma 3.3. *If $\mathbf{S}_j$ is drawn from a distribution such that for any fixed matrix A^T, $\mathbf{S}_j$ is a $1/2$ subspace embedding with probability $1-\delta$ and for any fixed matrices M, N, with probability $1-\delta$,*

$$\|M^\mathsf{T}\mathbf{S}_j^\mathsf{T}\mathbf{S}_j N - M^\mathsf{T}N\|_\mathsf{F} \le \sqrt{\varepsilon/n}\|M\|_\mathsf{F}\|N\|_\mathsf{F},$$

then with probability $\ge 1 - 2\delta$, $\|x^{*(j)} - \tilde{x}^{(j)}\|_2 \le (2\sqrt{\varepsilon})\|x^{*(j)}\|_2$.

Proof. Let $A = U\Sigma V^\mathsf{T}$ be the singular value decomposition of A. We have $x^{*(j)} = V\Sigma(I + \Sigma^2)^{-1}U^\mathsf{T}b^{(j)}$. By using $(I + \Sigma^2)^{-1} = \Sigma^{-1}(I + \Sigma^{-2})^{-1}\Sigma^{-1}$, we get $x^{*(j)} = V(I + \Sigma^{-2})^{-1}\Sigma^{-1}U^\mathsf{T}b^{(j)}$. Let $v^{(j)} = (I + \Sigma^{-2})^{-1}\Sigma^{-1}U^\mathsf{T}b^{(j)}$ which gives $x^{*(j)} = Vv^{(j)}$.

Similarly, $\tilde{x}^{(j)} = V(V^\mathsf{T}\mathbf{S}_j^\mathsf{T}\mathbf{S}_j V + \Sigma^{-2})^{-1}\Sigma^{-1}U^\mathsf{T}b^{(j)}$. Writing $V^\mathsf{T}\mathbf{S}_j^\mathsf{T}\mathbf{S}_j V = I_n + E$, we have

$$\begin{aligned}\tilde{x}^{(j)} &= V(I + \Sigma^{-2} + E)^{-1}\Sigma^{-1}U^\mathsf{T}b^{(j)}\\ &= V(I + (I + \Sigma^{-2})^{-1}E)^{-1}(I + \Sigma^{-2})^{-1}\Sigma^{-1}U^\mathsf{T}b^{(j)}\\ &= V(I + (I + \Sigma^{-2})^{-1}E)^{-1}v^{(j)}.\end{aligned}$$

As $\|E\|_2 \le 1/2$, the inverse $(I + (I + \Sigma^{-2})^{-1}E)^{-1}$ is well-defined. Since the matrix V has orthonormal columns, $\|\tilde{x}^{(j)} - x^{*(j)}\|_2 = \|(I + (I+\Sigma^{-2})^{-1}E)^{-1}v^{(j)} - v^{(j)}\|_2$. Let $(I + (I+\Sigma^{-2})^{-1}E)^{-1}v^{(j)} = v^{(j)} + \Delta$ and we have $v^{(j)} = v^{(j)} + (I+\Sigma^{-2})^{-1}Ev^{(j)} + (I + (I+\Sigma^{-2})^{-1}E)\Delta$ which implies $(I + (I+\Sigma^{-2})^{-1}E)\Delta = -(I+\Sigma^{-2})^{-1}Ev^{(j)}$. Finally,

$$\begin{aligned}(1/2)\|\Delta\|_2 &\le \sigma_{\min}(I + (I+\Sigma^{-2})^{-1}E)\|\Delta\|_2\\ &\le \|(I + (I+\Sigma^{-2})^{-1}E)\Delta\|_2\\ &= \|(I+\Sigma^{-2})^{-1}Ev^{(j)}\|_2 \le \|Ev^{(j)}\|_2.\end{aligned}$$

which gives $\|x^{*(j)} - \tilde{x}^{(j)}\|_2 = \|V\Delta\|_2 \le 2\|Ev^{(j)}\|_2$. If the matrix $\mathbf{S}_j$ has a $\sqrt{\varepsilon/n}$-AMM property i.e.,

$$\begin{aligned}\|V^\mathsf{T}\mathbf{S}_j^\mathsf{T}\mathbf{S}_j Vv^{(j)} - V^\mathsf{T}Vv^{(j)}\|_2 &\le \sqrt{\varepsilon/n}\|V\|_\mathsf{F}\|v^{(j)}\|_2\\ &= \sqrt{\varepsilon}\|v^{(j)}\|_2,\end{aligned}$$

we have $\|Ev^{(j)}\|_2 \le \sqrt{\varepsilon}\|Vv^{(j)}\|_2$ and that $\|x^{*(j)} - \tilde{x}^{(j)}\|_2 \le 2\sqrt{\varepsilon}\|v^{(j)}\|_2 = 2\sqrt{\varepsilon}\|x^{*(j)}\|_2$. □

Proof of Theorem 3.1. By a union bound, with probability $\ge 9/10$, in all t iterations, we can assume that the matrices $\mathbf{S}_j$ have both the subspace embedding property for the column space of A^T, as well as the AMM property for V and $v^{(j)}$.

From Lemma 3.2, $\|\hat{x} - x^*\|_2 = \|\tilde{x}^{(t)} + \sum_{i=1}^{t-1} \tilde{x}^{(i)} - x^*\|_2 = \|\tilde{x}^{(t)} - x^{*(t)}\|_2 \le (\sqrt{\varepsilon})\|x^{*(t)}\|_2$. We also have

$$x^* = x^{*(j-1)} + \sum_{i=1}^{j-2} \tilde{x}^{(i)} = x^{*(j)} + \sum_{i=1}^{j-1} \tilde{x}^{(i)}$$

which implies $x^{*(j)} = x^{*(j-1)} - \tilde{x}^{(j-1)}$ and therefore, $\|x^{*(j)}\|_2 = \|\tilde{x}^{(j-1)} - x^{*(j-1)}\|_2 \leq \sqrt{\varepsilon}\|x^{*(j-1)}\|_2$ for all j, where the last inequality follows from Lemma 3.3. Now noting that $x^{*(1)} = x^*$, we obtain $\|\hat{x} - x^*\|_2 \leq (\sqrt{\varepsilon})^t \|x^*\|_2$ and using the Pythagorean theorem,

$$\begin{aligned}\text{cost}(\hat{x}) &\leq \text{Opt} + (\sigma^2 + \lambda)\|\hat{x} - x^*\|_2^2 \\ &\leq \text{Opt} + (\sigma^2 + \lambda)\varepsilon^t \|x^*\|_2^2.\end{aligned}$$

As $\lambda\|x^*\|_2^2 \leq \text{Opt}$, we obtain the result. □

We now show that the OSNAP distribution has both the properties required by Algorithm 1.

3.1. Properties of OSNAP

Nelson & Nguyen (2013) proposed OSNAP, an oblivious subspace embedding. OSNAP embeddings are parameterized by their number m of rows and their sparsity s. Essentially, OSNAP is a random $m \times d$ matrix $\mathbf{S}$, with each column having exactly s nonzero entries at random locations. Each nonzero entry is $\pm 1/\sqrt{s}$ with probability $1/2$ each. They show that if the positions of the nonzero entries satisfy an "expectation" property and if the nonzero values are drawn from a k-wise independent distribution for a sufficiently large k, then $\mathbf{S}$ is an OSE.

Theorem 3.4 (Informal, (Nelson & Nguyen, 2013)). *If $m = O(n^{1+\gamma}\,\text{poly}(\log(n), 1/\varepsilon)/\varepsilon^2)$ and $s = O(1/\gamma\varepsilon)$, then OSNAP is an ε-OSE for n dimensional spaces. Further, $t_{\mathbf{S}A} = O(\texttt{nnz}(A)/\gamma\varepsilon)$ for any $d \times n$ matrix A.*

In the supplementary, we show that OSNAP with *any* sparsity parameter s and $m = \Omega(1/\varepsilon^2)$ has the ε-AMM property. We state our result as the following lemma.

Lemma 3.5. *OSNAP with $m = \Omega(1/\varepsilon^2\delta)$ and sparsity parameter $s \geq 1$ has the (ε, δ)-AMM property.*

3.2. Running times : Embedding vs Current Work

As discussed in the introduction, the algorithm of Chowdhury et al. (2018) is better than ours when $O(\log(1/\varepsilon))$ passes over the matrix A are allowed, as we require a fresh $1/2$ subspace embedding in each iteration and they require only one $1/2$ subspace embedding. However, our algorithm is faster when the algorithm is restricted to $t = O(1)$ passes over the input. We compare the running time of our algorithm with theirs when both algorithms are run only for 1 iteration to obtain $1 + \varepsilon$ approximate solutions. For ease of exposition, we consider the case when $\sigma^2/\lambda = O(1)$.

From Theorem 1.1, the algorithm of Chowdhury et al. (2018) requires a $c\sqrt{\varepsilon}$ subspace embedding to output a $1 + \varepsilon$ approximation to ridge regression. By applying a sequence of CountSketch and OSNAP sketches, we can obtain a $c\sqrt{\varepsilon}$ embedding with $m = n\,\text{poly}(\log(n))/\varepsilon$ and $t_{SA^\intercal} = O(\texttt{nnz}(A) + n^3\,\text{poly}(\log(n))/\sqrt{\varepsilon})$ or by directly applying OSNAP, we obtain $m = n\,\text{poly}(\log(n))/\varepsilon$ and $t_{SA^\intercal} = O(\texttt{nnz}(A)\,\text{poly}(\log(n))/\sqrt{\varepsilon})$.

From Theorem 3.1, our algorithm needs a random matrix that has the $1/2$ subspace embedding property and the $c\sqrt{\varepsilon/n}$-AMM property to compute a $1 + \varepsilon$ approximation. OSNAP with $m = O(n/\varepsilon + n\,\text{poly}(\log(n)))$ and $s = O(\text{poly}(\log(n)))$ has this property giving $t_{SA^\intercal} = O(\texttt{nnz}(A)\,\text{poly}(\log(n)))$.

Finally, the total time to compute $\tilde{x}$ is

$$O(t_{SA^\intercal} + mn^{\omega-1} + n^\omega),$$

where $\omega < 3$ denotes the matrix multiplication exponent. For the algorithm of Chowdhury et al. (2018), depending on the sketching matrices used as described above, the total running time is either

$$O(\texttt{nnz}(A) + n^3\,\text{poly}(\log(n))/\sqrt{\varepsilon} + n^\omega\,\text{poly}(\log(n))/\varepsilon)$$

or

$$O(\texttt{nnz}(A)\,\text{poly}(\log(n))/\sqrt{\varepsilon} + n^\omega\,\text{poly}(\log(n))/\varepsilon).$$

For Algorithm 1 with $t = 1$, the total running time is $O(\texttt{nnz}(A)\,\text{poly}(\log(n)) + n^\omega\,\text{poly}(\log(n))/\varepsilon)$. Thus we have that when $\texttt{nnz}(A) \approx n^\omega/\varepsilon$, our algorithm is asymptotically faster than their algorithm, as our running time does not have the n^3 term and $\texttt{nnz}(A)/\sqrt{\varepsilon}$ terms. We note that although the fastest matrix multiplication algorithms are sometimes considered impractical, Strassen's algorithm is already practical for reasonable values of n, and gives $\omega < \log_2 7$. If we consider the algorithm of Chowdhury et al. (2018) using just the OSNAP embedding, our algorithm is faster by a factor of $1/\sqrt{\varepsilon}$, which could be substantial when ε is small.

Even non-asymptotically, our result shows that we can replace the sketching matrix in their algorithm with a sketching matrix that is both sparser and has fewer rows, while still obtaining a $1 + \varepsilon$ approximation. Both of these properties help the algorithm to run faster.

4. Applications to Kernel Ridge Regression

A function $k : X \times X \to \mathbb{R}$ is called a positive semi-definite kernel if it satisfies the following two conditions: (i) For all $x, y \in X$, $k(x, y) = k(y, x)$, and (ii) for any finite set $S = \{s_1, \ldots, s_t\} \subseteq X$, the matrix $K = [k(s_i, s_j)]_{i,j\in[t]}$ is positive semi-definite. Mercer's theorem states that a function $k(\cdot, \cdot)$ is a positive semi-definite kernel as defined above if and only if there exists a function ϕ such that for all $x, y \in X$, $k(x, y) = \phi(x)^\intercal \phi(y)$. Many machine learning algorithms only work with inner products of the data points and therefore all such algorithms can work using

the function k directly instead of the explicit mapping ϕ, which in principle could even be infinite dimensional, for example, as in the case of the Gaussian kernel.

Let the rows of a matrix A be the input data points $a_1, \ldots, a_n$, and let $\phi(A)$ denote the matrix obtained by applying the function ϕ to each row of the matrix A. The kernel ridge regression problem (see Murphy (2012) for more details) is defined as

$$c^* = \arg\min_c \|\phi(A) \cdot c - b\|_2^2 + \lambda \|c\|_2^2.$$

We have that $c^* = \phi(A)^\mathsf{T}(\phi(A) \cdot \phi(A)^\mathsf{T} + \lambda I)^{-1} b$ and the value predicted for an input x is given by $\phi(x)^\mathsf{T} c^* = \phi(x)^\mathsf{T}\phi(A)^\mathsf{T}(\phi(A) \cdot \phi(A)^\mathsf{T} + \lambda I)^{-1} b$. Letting $\beta = (\phi(A)\phi(A)^\mathsf{T} + \lambda I)^{-1} b$ we have $\phi(x)^\mathsf{T} c^* = \sum_i k(a_i, x)\beta_i$. Now, note that the (i,j)-th entry of the matrix $K := \phi(A) \cdot \phi(A)^\mathsf{T}$ is given by $k(a_i, a_j)$ and therefore, to solve the kernel ridge regression problem, we do not need the explicit map $\phi(\cdot)$ and can work directly with the kernel function. Nevertheless, to construct the matrix K, we need to query the kernel function k for $\Theta(n^2)$ pairs of inputs, which may be prohibitive if the kernel evaluation is slow.

Our result for ridge regression shows that if $\mathbf{S}$ is a $1/2$ subspace embedding and gives a $\varepsilon/2\sqrt{n}$ AMM guarantee, then

$$\tilde{c} = \phi(A)^\mathsf{T} \cdot (\phi(A) \cdot \mathbf{S}^\mathsf{T}\mathbf{S} \cdot \phi(A)^\mathsf{T} + \lambda I)^{-1} b$$

satisfies $\|\tilde{c} - c^*\| \leq \varepsilon \|c^*\|_2$ and if $\tilde{\beta} := (\phi^\mathsf{T}(A) \cdot \mathbf{S}^\mathsf{T}\mathbf{S} \cdot \phi(A)^\mathsf{T} + \lambda I)^{-1} b$, then for a new input x, the prediction on x can be computed as $\sum_i k(a_i, x)\tilde{\beta}_i$. For polynomial kernels, $k(x, y) = \langle x, y \rangle^p$, given the matrix A, it is possible to compute $\mathbf{S} \cdot \phi(A)^\mathsf{T}$ for a random matrix $\mathbf{S}$ that satisfies both the subspace embedding property and the AMM property, and hence obtain $\tilde{\beta}$ without computing the kernel matrix. The next theorem follows from the proofs of Theorems 1 and 3 of Ahle et al. (2020).

Theorem 4.1. *For all positive integers n, d, p, there exists a distribution on linear sketches $\Pi^p \in \mathbb{R}^{m \times d^p}$ parameterized by sparsity s such that: if $m = \Omega(p/\varepsilon^2)$ and any sparsity s, then Π^p has the ε-AMM property, while if $m = \tilde{\Omega}(p^4 n/\varepsilon^2)$ and $s = \tilde{\Omega}((p^4/\varepsilon^2 \operatorname{poly}(\log(nd/\varepsilon)))$, then Π^p has the ε subspace embedding property. Further, given any matrix $A \in \mathbb{R}^{n \times d}$, the matrix $\Pi^p \cdot \phi(A)^\mathsf{T}$ for $\phi(x) = x^{\otimes p}$ can be computed in $\tilde{O}(pnm + ps \cdot \mathtt{nnz}(A))$ time.*

We show that the construction of Ahle et al. (2020) gives the above theorem when S_{base} is taken to be TensorSketch and T_{base} is taken to be OSNAP. To prove the theorem, we first prove a lemma which shows that the OSNAP distribution has the JL-moment property. For a random variable $\mathbf{X}$, let $\|\mathbf{X}\|_{L^t} := (\mathbb{E}[|\mathbf{X}|^t])^{1/t}$.

Definition 4.2 (JL-Moment Property)**.** For every positive integer t and parameters $\varepsilon, \delta \geq 0$, we say a random matrix $\mathbf{S} \in \mathbb{R}^{m \times d}$ satisfies the (ε, δ, t)-JL moment property if for any $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$,

$$\|\|\mathbf{S}x\|_2^2 - 1\|_{L^t} \leq \varepsilon \delta^{1/t} \text{ and } \mathbb{E}[\|\mathbf{S}x\|_2^2] = 1.$$

Lemma 4.3. *If $\mathbf{S}$ is an OSNAP matrix with $m = \Omega(1/\delta\varepsilon^2)$ rows and any sparsity parameter $s \geq 1$, then $\mathbf{S}$ has the $(\varepsilon, \delta, 2)$-JL moment property.*

Proof of Theorem 4.1. Let $q = 2^{\lceil \log_2(p) \rceil}$. The construction of the sketch for polynomial kernels of Ahle et al. (2020) uses two distributions of matrices S_{base} and T_{base}. The proof of Theorem 1 of Ahle et al. (2020) requires that the distributions S_{base} and T_{base} have the $(\varepsilon/\sqrt{4q+2}, \delta, 2)$-JL moment property. We take S_{base} to be TensorSketch and T_{base} to be OSNAP. As Lemma 4.3 shows, OSNAP with $m = \Omega(q/\delta\varepsilon^2)$ and any sparsity s has the $(\varepsilon/\sqrt{4q+2}, \delta, 2)$-JL moment property.

From Theorem 3 of Ahle et al. (2020), we also have that for $m = \tilde{\Omega}(p^4 n/\varepsilon^2)$ and sparsity parameter $s = \tilde{\Omega}((p^4/\varepsilon^2) \cdot \operatorname{poly}(\log(nd/\varepsilon)))$, the sketch has the ε-subspace embedding property. The running time of applying the sketch to $\phi(A)^\mathsf{T}$ also follows from the same theorem. □

Thus for the sketch to have both the $1/2$ subspace embedding property and the $\varepsilon/\sqrt{4n}$ AMM property, we need to take $m = \tilde{\Omega}(p^4 n + pn/\varepsilon^2)$ and $s = \tilde{\Omega}(p^4 \operatorname{poly}(\log(nd)))$. The time to compute $\Pi^p \cdot \phi(A)^\mathsf{T}$ is $\tilde{O}(p^5 \mathtt{nnz}(A) + p^5 n^2 + p^2 n/\varepsilon^2)$ and the time to compute $\tilde{\beta}$ is $\tilde{O}(p^5 \mathtt{nnz}(A) + p^5 n^2 + p^2 n^2/\varepsilon^2 + p^4 n^\omega + pn^\omega/\varepsilon^2)$, thereby obtaining a near-input sparsity time algorithm for polynomial kernel ridge regression.

5. Lower bounds

Dimensionality reduction, by multiplying the input matrix A on the right with a random sketching matrix, seems to be the most natural way to speed up ridge regression. Recall that in our algorithm above, we show that we only need the sketching distribution to satisfy a simple AMM guarantee, along with being a constant factor subspace embedding, to be able to obtain a $1 + \varepsilon$ approximation. We show that, in this natural framework, the bounds on the number of rows required for a sketching matrix we obtain are nearly optimal for all "non-dilating" distributions.

More formally, we show lower bounds in the restricted setting where for an oblivious random matrix $\mathbf{S}$, the vector $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I)^{-1} b$ must be a $1 + \varepsilon$ approximation to the ridge regression problem with probability $\geq 99/100$. We show that the matrix $\mathbf{S}$ must at least have $m = \Omega(n\sigma^2/\lambda\varepsilon)$ rows if $\mathbf{S}$ is "non-dilating".

Definition 5.1 (Non-Dilating Distributions)**.** A distribution $\mathcal{S}$ over $m \times d$ matrices is a Non-Dilating distribution if for

all $d \times n$ orthonormal matrices V,

$$\Pr_{\mathbf{S}\sim\mathcal{S}}[\|\mathbf{S}V\|_2 \leq O(1)] \geq 99/100.$$

Most sketching distributions proposed in previous work satisfy the property $\mathbb{E}[V^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}V] = V^\mathsf{T}V = I$. Thus the condition of non-dilation is not very restrictive. For example, a Gaussian distribution with $O(n)$ rows satisfies this condition, and other sketching distributions such as SRHT, CountSketch, and OSNAP with $O(n\log(n))$ rows all satisfy this condition with $O(1)$ replaced by at most $O(\log(n))$. Though we prove our lower bounds for non-dilating distributions with $O(1)$ distortion, the lower bounds also hold with distributions with $O(\log(n))$ distortion with at most an $O(\log(n))$ factor loss in the lower bound.

For $n' \geq n$, let $O^{n'\times n}$ denote the collection of $n' \times n$ orthonormal matrices $V \in \mathbb{R}^{n'\times n}$ i.e., $V^\mathsf{T}V = I_n$. Without loss of generality, we assume that $\lambda = 1$.

Assume that there is a distribution $\mathcal{S}$ over $m \times d$ matrices such that given an arbitrary matrix $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ such that for $\mathbf{S} \sim \mathcal{S}$, with probability $\geq 99/100$,

$$\|A\tilde{x} - b\|_2^2 + \|\tilde{x}\|_2^2 \leq (1+\varepsilon)\text{Opt},$$

where $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T}+I)^{-1}b$. Given an instance (A, b), let $\mathbf{S}$ be a good$_{A,b}$ matrix if the above event holds, i.e., $\tilde{x}$ is a $1+\varepsilon$ approximation. Let b be a fixed unit vector. Thus, from our assumption,

$$\Pr_{\mathbf{U}\sim O^{n\times n},\mathbf{V}\sim O^{d\times n},\mathbf{S}\sim\mathcal{S}}[\mathbf{S} \text{ is good}_{\sigma\mathbf{U}\mathbf{V}^\mathsf{T},b}] \geq 99/100. \tag{3}$$

For the problem $(\sigma\mathbf{U}\mathbf{V}^\mathsf{T}, b)$ where b is a fixed unit vector, we have Opt $= 1/(1+\sigma^2)$. We also have for $v = (\Sigma^{-2} + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\Sigma^{-1}\mathbf{U}^\mathsf{T}b$ that

$$\begin{aligned}\text{cost}(\tilde{x}) - \text{Opt} &= v^\mathsf{T}E\Sigma(I - (\Sigma^2+I)^{-1})\Sigma Ev \\ &\geq \lambda_{\min}(I - (\Sigma^2+I)^{-1})\|\Sigma Ev\|_2^2,\end{aligned}$$

where $E = \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V} - \mathbf{V}^\mathsf{T}\mathbf{V}$, which is the error in approximating the identity matrix using the sketch $\mathbf{S}$, and Σ is the matrix of singular values of $\sigma\mathbf{U}\mathbf{V}^\mathsf{T}$. In our case, $\Sigma = \sigma I_n$ for some $\sigma \geq 1$ which implies that

$$\text{cost}(\tilde{x}) - \text{Opt} \geq \frac{1}{2}\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\mathbf{U}^\mathsf{T}b\|_2^2$$

once we cancel out Σ and Σ^{-1}. Thus, if $\mathbf{S}$ is good$_{(\sigma\mathbf{U}\mathbf{V}^\mathsf{T},b)}$,

$$\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\mathbf{U}^\mathsf{T}b\|_2^2 \leq \frac{2\varepsilon}{1+\sigma^2} \leq \frac{2\varepsilon}{\sigma^2}.$$

Therefore, $\Pr_{\mathbf{U},\mathbf{V},\mathbf{S}}[\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\mathbf{U}^\mathsf{T}b\|_2^2 \leq 2\varepsilon/\sigma^2] \geq \Pr_{\mathbf{U},\mathbf{V},\mathbf{S}}[\mathbf{S} \text{ is good}_{\sigma\mathbf{U}\mathbf{V}^\mathsf{T},b}] \geq 99/100$. Now, for a fixed unit vector b, the vector $\mathbf{U}^\mathsf{T}b$ is a uniformly random unit vector that is independent of $\mathbf{V}$ and $\mathbf{S}$. Thus,

$$\Pr_{\mathbf{V},\mathbf{S},\mathbf{r}}[\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\mathbf{r}\|_2^2 \leq 2\varepsilon/\sigma^2] \geq 0.99,$$

where above and throughout the section, $\mathbf{r}$ is a uniformly random unit vector. Now we transform this property of the random matrix $\mathbf{S}$ into a probability statement about the Frobenius norm of a certain matrix.

Lemma 5.2 (Random vector to Frobenius Norm)**.** *If $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a random matrix independent of the random uniform vector $\mathbf{r}$ such that $Pr_{\mathbf{M},\mathbf{r}}[\|\mathbf{M}\mathbf{r}\|_2^2 \leq a] \geq 99/100$, then $Pr_{\mathbf{M}}[\|\mathbf{M}\|_\mathsf{F}^2 \leq Can] \geq 9/10$ for large enough constant C.*

This lemma implies that for any random matrix $\mathbf{S}$ satisfying (3), we have $\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\|_\mathsf{F}^2 \leq Cn\varepsilon/\sigma^2$ with probability $\geq 9/10$ over $\mathbf{V}, \mathbf{S}$. Using the non-dilating property of $\mathbf{S}$ and applying a union bound, we now have with probability $\geq 8/10$,

$$\begin{aligned}\|E\|_\mathsf{F}^2 &\leq \frac{\|E(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1}\|_\mathsf{F}^2}{\sigma_{\min}((\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1})^2} \\ &= \frac{Cn\varepsilon/\sigma^2}{\sigma_{\min}((\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V})^{-1})^2} \leq O(n\varepsilon/\sigma^2)\end{aligned}$$

where we used the fact that for any invertible matrix A, $1/\sigma_{\min}(A^{-1}) = \sigma_{\max}(A)$ and $\sigma_{\max}(\sigma^{-2}I + \mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V}) \leq (1/\sigma^2) + \|\mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V}\|_2 = O(1)$ with probability $\geq 9/10$. Thus, a lower bound on the number of rows in the matrix $\mathbf{S}$ to obtain, with probability $\geq 8/10$,

$$\|\mathbf{V}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V} - I\|_\mathsf{F} \leq O(\sqrt{n\varepsilon/\sigma^2}) = O(\sqrt{\varepsilon/n\sigma^2})n \tag{4}$$

implies a lower bound on the number of rows of a random matrix $\mathbf{S}$ that satisfies (3).

5.1. Lower bounds for AMM

Lemma 5.3. *Given parameters n and error parameter $\varepsilon \leq c/\sqrt{n}$ for a small enough constant c, for all $d \geq Cn/\varepsilon^2$, if a random matrix $\mathbf{S} \in \mathbb{R}^{m\times d}$ for all matrices $A \in \mathbb{R}^{d\times n}$ satisfies, $\|A^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}A - A^\mathsf{T}A\|_\mathsf{F} \leq \varepsilon\|A^\mathsf{T}\|_\mathsf{F}\|A\|_\mathsf{F}$ with probability $\geq 9/10$, then $m = \Omega(1/\varepsilon^2)$.*

Moreover, the lower bound of $\Omega(1/\varepsilon^2)$ holds even for the sketching matrices that give the following guarantee: $Pr_{\mathbf{A},\mathbf{S}}[\|\mathbf{A}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{A} - I\|_\mathsf{F} \leq \varepsilon n] \geq 0.9$, where $\mathbf{A}$ is a uniformly random $d \times n$ orthonormal matrix independent of the sketch $\mathbf{S}$.

Although the above lemma only shows that an AMM sketch requires $m = \Omega(1/\varepsilon^2)$ for $d \geq Cn/\varepsilon^2$, we can extend it to show the lower bound for $d \geq C/\varepsilon^2$ for a large enough constant C. Note that $C/\varepsilon^2 = \Omega(n)$ since $\varepsilon \leq c/\sqrt{n}$.

Theorem 5.4. *Given $n \geq 0$ and $\varepsilon < c/\sqrt{n}$ for a small enough constant c, there are universal constants C, D such that for all $d \geq D/\varepsilon^2$, any distribution that has the ε AMM property for $d \times n$ matrices must have $\geq C/\varepsilon^2$ rows.*

As discussed in the introduction, we crucially use the fact that a sub-matrix of a random orthonormal matrix is close

to a Gaussian matrix in total variation distance to prove the above theorem. This seems to be a useful direction to obtain lower bounds for other sketching problems.

5.2. Lower Bound Wrapup

In the case of ridge regression with $\lambda = 1$, (4) shows that the sketching distribution has to satisfy the AMM guarantee with parameter $c\sqrt{\varepsilon/n\sigma^2}$. By using the above hardness result for AMM, we obtain the following theorem.

Theorem 5.5. *If $\mathcal{S}$ is a non-dilating distribution over $m \times d$ matrices such that for all ridge regression instances (A, b, λ) with $A \in \mathbb{R}^{n\times d}$, $1 \leq \sigma^2/\lambda \leq \alpha$ satisfies,*

$$Pr_{\mathbf{S}\sim\mathcal{S}}[\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leq (1+\varepsilon)\text{Opt}] \geq 0.99,$$

for $\tilde{x} = A^\mathsf{T}(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I)^{-1}b$, then $m = \Omega(n\alpha/\varepsilon) = \Omega(n\sigma^2/\lambda\varepsilon)$.

6. Communication Complexity Lower Bounds for Ridge Regression

Consider a ridge regression matrix A of the form $A_1 + A_2$ where Alice has the matrix A_1 and Bob has the matrix A_2. To compute a vector y such that $A^\mathsf{T}y$ is a $1+\varepsilon$ approximation, the theorem from the previous section lower bounds the communication required between Alice and Bob by $\Omega(n\sigma^2/\lambda\varepsilon)$. The lower bound is crude in that it assumes that they only communicate $A_1\mathbf{S}^\mathsf{T}$ between them for some sketch $\mathbf{S}$, and they compute y only as $(A\mathbf{S}^\mathsf{T}\mathbf{S}A^\mathsf{T} + \lambda I)^{-1}b$.

In this section, we present communication complexity lower bounds for a different but related problem of computing a value v such that $v = (1 \pm \varepsilon)\text{Opt}$ in the above two-player scheme, where Alice has the matrix A_1 and Bob has the matrix A_2. We show the lower bound by reducing from the well-known GAP-HAMMING problem (Indyk & Woodruff, 2003; Woodruff, 2004) to a ridge regression instance with a $2 \times d$ design matrix.

In the GAP-HAMMING problem, Alice and Bob receive vectors $x, y \in \{\pm 1\}^d$ and want to decide if $d_H(x, y) \geq d/2 + \varepsilon d$ or $d_H(x, y) \leq d/2 - \varepsilon d$, where $d_H(x, y)$ is the Hamming distance $|\{i \,|\, x_i \neq y_i\}|$. This problem has a communication complexity lower bound of $\Omega(1/\varepsilon^2)$ bits, even for multiple rounds (Chakrabarti & Regev, 2012). Let M be a $2 \times d$ matrix with x and y as its rows. Consider the following ridge regression problem:

$$\min_x \|Mx - \begin{bmatrix} 1 \\ -1 \end{bmatrix}\|_2^2 + \lambda\|x\|_2^2.$$

Let $N = d_H(x, y)$. The optimal value of the above ridge regression problem is given by $2\lambda/(\lambda + 2N)$. In the case when $N \geq d/2 + \varepsilon d$ and $N \leq d/2 + \varepsilon d$, the optimal values differ by a factor of $1 + 4\varepsilon/(1 + \lambda/d)$. So, obtaining a $1 + O(\varepsilon/(1 + \lambda/d))$ approximation to ridge regression lets us solve the Gap-Hamming problem, and hence requires $\Omega(1/\varepsilon^2)$ bits of communication. As $\|M\|_2^2 = \Theta(d)$, we obtain an $\Omega(1/\varepsilon^2(1+\lambda/\sigma^2))$ bit lower bound for computing a $1 + \varepsilon$ approximate value for ridge regression.

In contrast to the $\Omega(1/\varepsilon^2)$ type communication lower bounds on computing $1 \pm \varepsilon$ approximations to the optimal values of ridge regression, we obtain $\Omega(d)$ lower bounds on the communication complexity of approximating optimal values of Lasso and square-root Lasso objectives even up to a factor of $1 + c$ for a small enough constant $c > 0$. Concretely, we prove the following results.

Theorem 6.1 (Communication Complexity of Lasso)**.** *Let $0 < \lambda < 1$ be the Lasso parameter. If Alice has the $n \times d$ matrix M_1 and Bob has the $n \times d$ matrix M_2, then to determine a $1 + c$ approximation, for a small enough constant c, to the optimal value of*

$$\min_z \|(M_1 + M_2)z - b\|_2^2 + \lambda\|z\|_1,$$

requires $\Omega(d)$ bits of communication between Alice and Bob.

Theorem 6.2 (Hardness of Sketching Square-Root Lasso)**.** *Let $0 < \lambda < 2\sqrt{2}/3$ be the Lasso parameter. If Alice has the $n \times d$ matrix M_1 and Bob has the $n \times d$ matrix M_2, then to determine a $1+c$ approximation, for a small enough constant $c > 0$, to the optimal value of*

$$\min_z \|(M_1 + M_2)z - b\|_2 + \lambda\|z\|_1,$$

requires $\Omega(d)$ bits of communication between Alice and Bob.

To show these lower bounds, we reduce the classic DISJOINTNESS problem (Håstad & Wigderson, 2007) to computing $1 + c$ approximations to optimal values of an appropriate Lasso and square-root Lasso problem. See the supplementary for proofs.

Acknowledgments

The authors thank National Institute of Health (NIH) grant 5401 HG 10798-2, Office of Naval Research (ONR) grant N00014-18-1-2562, and a Simons Investigator Award.

References

Ahle, T. D., Kapralov, M., Knudsen, J. B., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 141–160. SIAM, 2020.

Assadi, S. and Raz, R. Near-quadratic lower bounds for two-pass graph streaming algorithms. In *2020 IEEE 61st*

Annual Symposium on Foundations of Computer Science (FOCS), pp. 342–353. IEEE, 2020.

Brody, J. and Woodruff, D. P. Streaming algorithms with one-sided estimation. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 436–447. Springer, 2011.

Chakrabarti, A. and Regev, O. An optimal lower bound on the communication complexity of Gap-Hamming-Distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

Chen, L., Kol, G., Paramonov, D., Saxena, R., Song, Z., and Yu, H. Near-optimal two-pass streaming algorithm for sampling random walks over directed graphs. *arXiv preprint arXiv:2102.11251*, 2021.

Chowdhury, A., Yang, J., and Drineas, P. An iterative, sketching-based framework for ridge regression. In *International Conference on Machine Learning*, pp. 989–998. PMLR, 2018.

Clarkson, K. L. and Woodruff, D. P. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6), January 2017. ISSN 0004-5411. doi: 10.1145/3019134.

Cohen, M. B. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 278–287. SIAM, 2016.

Håstad, J. and Wigderson, A. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

Indyk, P. and Woodruff, D. P. Tight lower bounds for the distinct elements problem. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*, pp. 283–288. IEEE Computer Society, 2003. doi: 10.1109/SFCS.2003.1238202.

Jiang, T. and Ma, Y. Distances between random orthogonal matrices and independent normals. *arXiv preprint arXiv:1704.05205*, 2017.

Konrad, C. and Naidu, K. K. On two-pass streaming algorithms for maximum bipartite matching. *arXiv preprint arXiv:2107.07841*, 2021.

Li, Y. and Woodruff, D. P. The product of Gaussian matrices is close to Gaussian. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, 2021.

Lowther, G. Anti-concentration of Gaussian Quadratic form. MathOverflow.net, 2012. URL https://mathoverflow.net/q/95108.

Meng, X. and Mahoney, M. W. 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*, pp. 91–100, 2013.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Nelson, J. and Nguyen, H. L. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*, pp. 117–126, 2013.

Nelson, J. and Nguyen, H. L. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*, pp. 883–894. Springer, 2014.

Park, Y., Han, D.-J., Kim, D.-Y., Seo, J., and Moon, J. Few-round learning for federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Woodruff, D. P. Optimal space lower bounds for all frequency moments. In Munro, J. I. (ed.), *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, pp. 167–175. SIAM, 2004.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014. ISSN 1551-305X. doi: 10.1561/0400000060.

A. Missing Proofs from Section 4

Proof of Lemma 4.3. For $i \in [m]$ and $j \in [d]$, let $\delta_{i,j}$ be the indicator random variable that denotes if the (i,j)-th entry of the matrix $\mathbf{S}$ is nonzero. We have that $\sum_i \delta_{i,j} = s$ and that for any $S \subseteq [m] \times [d]$, $\mathbb{E}[\Pi_{(i,j)\in S}\delta_{i,j}] \leq (s/m)^{|S|}$. Also, let $\sigma_{i,j}$ be the sign of the (i,j)-th entry and let $\sigma_{i,j}$ be 4-wise independent Rademacher random variables. Now,

$$\|\mathbf{S}x\|_2^2 = \sum_i |\mathbf{S}_{i*}x|^2 = \frac{1}{s}\sum_i (\sum_j \delta_{i,j}\sigma_{i,j}x_j)^2 = \frac{1}{s}\sum_i \sum_{j,j'} \delta_{i,j}\delta_{i,j'}\sigma_{i,j}\sigma_{i,j'}x_j x_{j'}$$
$$= \frac{1}{s}\sum_i \sum_j (\delta_{i,j})^2(\sigma_{i,j})^2 x_j^2 + \frac{1}{s}\sum_i \sum_{j\neq j'} \delta_{i,j}\delta_{i,j'}\sigma_{i,j}\sigma_{i,j'}x_j x_{j'}.$$

We have $\delta_{i,j}^2 = \delta_{i,j}$ and $\sigma_{i,j}^2 = 1$ for all i, j. So,

$$(1/s)\sum_i \sum_j (\delta_{i,j})^2(\sigma_{i,j})^2 x_j^2 = (1/s)\sum_i \sum_j \delta_{i,j}x_j^2 = (1/s)\sum_j x_j^2 \sum_i \delta_{i,j} = (1/s)\sum_j x_j^2 \cdot s = \|x\|_2^2 = 1.$$

Therefore,

$$\|\mathbf{S}x\|_2^2 = 1 + \frac{1}{s}\sum_i \sum_{j\neq j'} \delta_{i,j}\delta_{i,j'}\sigma_{i,j}\sigma_{i,j'}x_j x_{j'}.$$

If $\sigma_{i,j}$ are uniform random signs that are 2-wise independent, then for $j \neq j'$, $\mathbb{E}[\sigma_{i,j}\sigma_{i,j'}] = 0$ and as the set of random variables $\delta_{i,j}$ are independent of the random variables $\sigma_{i,j}$, we have $\mathbb{E}[\delta_{i,j}\delta_{i,j'}\sigma_{i,j}\sigma_{i,j'}] = \mathbb{E}[\delta_{i,j}\delta_{i,j'}]\mathbb{E}[\sigma_{i,j}\sigma_{i,j'}] = 0$ for $j \neq j'$ which implies that $\mathbb{E}[\|\mathbf{S}x\|_2^2] = 1$. We also have

$$(\|\mathbf{S}x\|_2^2 - 1)^2 = \frac{1}{s^2}\sum_{i,i'} \sum_{\substack{j\neq j' \\ k\neq k'}} \delta_{i,j}\delta_{i,j'}\delta_{i',k}\delta_{i',k'}\sigma_{i,j}\sigma_{i,j'}\sigma_{i',k}\sigma_{i',k'}x_j x_{j'}x_k x_{k'}.$$

If $i \neq i'$, $j \neq j'$, and $k \neq k'$, then the random variables $\sigma_{i,j}$, $\sigma_{i,j'}, \sigma_{i',k}$, and $\sigma_{i',k'}$ are distinct and if they are 4-wise independent Rademacher random variables, then $\mathbb{E}[\sigma_{i,j}\sigma_{i,j'}\sigma_{i',k}\sigma_{i',k'}] = 0$ which implies that

$$\mathbb{E}[(\|\mathbf{S}x\|_2^2 - 1)^2] = \frac{1}{s^2}\sum_i \sum_{\substack{j\neq j' \\ k\neq k'}} \mathbb{E}[\delta_{i,j}\delta_{i,j'}\delta_{i,k}\delta_{i,k'}]\mathbb{E}[\sigma_{i,j}\sigma_{i,j'}\sigma_{i,k}\sigma_{i,k'}]x_j x_{j'}x_k x_{k'}.$$

Again, if all the indices j, j', k, k' are distinct, then by the 4-wise independence of the σ random variables, we obtain that $\mathbb{E}[\sigma_{i,j}\sigma_{i,j'}\sigma_{i,k}\sigma_{i,k'}] = 0$, which leaves only $j = k \neq j' = k'$ and $j = k' \neq j' = k$ as the cases where the expectation can be non-zero. In each of these cases, $\sigma_{i,j}\sigma_{i,j'}\sigma_{i,k}\sigma_{i,k'} = 1$ with probability 1. Therefore,

$$\mathbb{E}[(\|\mathbf{S}x\|_2^2 - 1)^2] = \frac{2}{s^2}\sum_i \sum_{j,j'} \mathbb{E}[\delta_{i,j}\delta_{i,j'}]x_j^2 x_{j'}^2 \leq \frac{2}{s^2}\frac{s^2}{m^2}\sum_i \sum_{j\neq j'} x_j^2 x_{j'}^2 \leq \frac{2}{m^2}\sum_i (\sum_j x_j^2)(\sum_{j'} x_{j'}^2) \leq \frac{2}{m}$$

which gives that $\|\|\mathbf{S}x\|_2^2 - 1\|_{L^2} = \mathbb{E}[(\|\mathbf{S}x\|_2^2 - 1)^2]^{1/2} \leq \sqrt{2/m}$. Now, for $m = \Omega(1/\varepsilon^2\delta)$, we have $\|\|\mathbf{S}x\|_2^2 - 1\|_{L^2} \leq \varepsilon\delta^{1/2}$, which proves that the matrix $\mathbf{S}$ has the $(\varepsilon, \delta, 2)$-JL moment property. □

B. Missing Proofs from Section 5

B.1. Proof of Lemma 5.2

Lemma 5.2 transforms a probability statement about the squared norm of a product of a random matrix $\mathbf{M}$ with an independent uniform random vector $\mathbf{r}$. To prove the lemma, we first prove the following similar lemma in which the matrix M is a deterministic matrix.

Lemma B.1. *Let $M \in \mathbb{R}^{n\times n}$ be a fixed matrix and $\mathbf{r}$ be a uniformly random unit vector. If $Pr_{\mathbf{r}}[\|M\mathbf{r}\|_2^2 \leq a] \geq 9/10$, then $\|M\|_F^2 \leq Cna$ for a large enough universal constant C.*

Proof. Let $\mathbf{g} \in \mathbb{R}^n$ be a Gaussian random vector with i.i.d. entries drawn from $N(0,1)$. Then the distribution of $\mathbf{g}/\|\mathbf{g}\|_2$ is identical to that of a uniformly random unit vector in n dimensions by rotational invariance of the Gaussian distribution. Therefore from our assumption, $\Pr_{\mathbf{g}}[\|M\mathbf{g}\|_2^2 \leq a\|\mathbf{g}\|_2^2] \geq 9/10$. We also have that with probability $\geq 9/10$, $\|\mathbf{g}\|_2^2 \leq C_1 n$ for a large enough absolute constant C_1. Thus, we have by a union bound that,

$$\Pr_{\mathbf{g}}[\|M\mathbf{g}\|_2^2 \leq a\|\mathbf{g}\|_2^2 \ \wedge \ \|\mathbf{g}\|_2^2 \leq C_1 n] \geq 8/10,$$

which implies that

$$\Pr_{\mathbf{g}}[\|M\mathbf{g}\|_2^2 \leq C_1 an] \geq 8/10.$$

Let $M = U\Sigma V^{\mathsf{T}}$ be the singular value decomposition of the matrix M. Then, the above equation is equivalent to

$$8/10 \leq \Pr_{\mathbf{g}}[\|\Sigma V^{\mathsf{T}}\mathbf{g}\|_2^2 \leq C_1 an] = \Pr_{\mathbf{g}}[\|\Sigma\mathbf{g}\|_2^2 \leq C_1 an]$$

where the equality follows from the fact that for an orthonormal matrix V^{T}, we have $V^{\mathsf{T}}\mathbf{g} \equiv \mathbf{g}$. Thus, if the singular values of M are $\sigma_1, \ldots, \sigma_n$, we have

$$\Pr_{\mathbf{g}}[\sum_i \sigma_i^2 \mathbf{g}_i^2 \leq C_1 an] \geq 8/10.$$

Now, we have the following lemma which gives an upper bound on the probability of a linear combination of squared Gaussian random variables being small.

Lemma B.2 (Lowther (2012))**.** *If $a_i \geq 0$ for $i = 1, \ldots, n$ are constants and $\mathbf{g}_1, \ldots, \mathbf{g}_n$ are i.i.d. Gaussians of mean 0 and variance 1, then for every $\delta > 0$,*

$$Pr[\sum_i a_i \mathbf{g}_i^2 \leq \delta \sum_i a_i] \leq e\sqrt{\delta}.$$

Proof. The inequality is obviously true for $\delta \geq 1$. We now consider arbitrary $\delta < 1$. Assume without loss of generality that $\sum_i a_i = 1$. Now, for any $\lambda > 0$,

$$\Pr[\sum_i a_i \mathbf{g}_i^2 \leq \delta] = \Pr[-\lambda \sum_i a_i \mathbf{g}_i^2 \geq -\lambda\delta] = \Pr[\exp(-\lambda \sum_i a_i \mathbf{g}_i^2) \geq \exp(-\lambda\delta)] \leq \exp(\lambda\delta)\mathbb{E}[\exp(-\lambda \sum_i a_i \mathbf{g}_i^2)]$$

and therefore,

$$\begin{aligned}\Pr[\sum_i a_i \mathbf{g}_i^2 \leq \delta] &= \exp(\lambda\delta)\mathbb{E}[\exp(-\lambda \sum_i a_i \mathbf{g}_i^2)] \\ &= \exp(\lambda\delta)\prod_i \mathbb{E}[\exp(-\lambda a_i \mathbf{g}_i^2)] \\ &= \exp(\lambda\delta)\prod_i (1+2\lambda a_i)^{-1/2}.\end{aligned}$$

Now, $\prod_i (1+2\lambda a_i) \geq 1 + 2\lambda(\sum_i a_i) = 1 + 2\lambda$ which implies that $\prod (1+2\lambda a_i)^{-1/2} \leq (1+2\lambda)^{-1/2}$ which gives

$$\Pr[\sum_i a_i \mathbf{g}_i^2 \leq \delta] \leq \exp(\lambda\delta)(1+2\lambda)^{-1/2}.$$

Picking $\lambda \geq 0$ such that $1 + 2\lambda = 1/\delta$, we obtain that $\Pr[\sum_i a_i \mathbf{g}_i^2 \leq \delta] \leq \exp((1-\delta)/2)\sqrt{\delta} \leq e\sqrt{\delta}$. □

For $\delta = 0.01$, the above lemma implies that $\Pr[\sum_i \sigma_i^2 \mathbf{g}_i^2 \leq 0.01 \sum_i \sigma_i^2] \leq e \cdot (0.1) \leq 0.3$. This, in particular implies that $0.01 \sum_i \sigma_i^2 = 0.01\|\Sigma\|_{\mathsf{F}}^2 \leq C_1 an$ which gives $\|M\|_{\mathsf{F}}^2 = \|\Sigma\|_{\mathsf{F}}^2 \leq Can$ for a large enough absolute constant C. □

We now extend the above lemma to the case when the matrix $\mathbf{M}$ is also random and independent of the random unit vector $\mathbf{r}$.

Proof of Lemma 5.2. Let $\mathbf{M}$ be *good* if $\Pr_{\mathbf{r}}[\|\mathbf{M}\mathbf{r}\|_2^2 \leq a] \geq 9/10$ and let $\mathbf{M}$ be *bad* otherwise and note from the above lemma that if $\mathbf{M}$ is *good*, then $\|\mathbf{M}\|_{\mathsf{F}}^2 \leq Can$. Now,

$$\begin{aligned}99/100 &\leq \Pr_{\mathbf{M},\mathbf{r}}[\|\mathbf{M}\mathbf{r}\|_2^2 \leq a] \\ &\leq \Pr_{\mathbf{M}}[\mathbf{M} \text{ is } \textit{good}] + \Pr_{\mathbf{M}}[\mathbf{M} \text{ is } \textit{bad}] \cdot (9/10) \\ &= 9/10 + (1/10) \cdot \Pr_{\mathbf{M}}[\mathbf{M} \text{ is } \textit{good}]\end{aligned}$$

which implies that $\Pr_{\mathbf{M}}[\mathbf{M} \text{ is } \textit{good}] \geq 9/10$ and therefore $\Pr_{\mathbf{M}}[\|\mathbf{M}\|_{\mathsf{F}}^2 \leq Can] \geq \Pr_{\mathbf{M}}[\mathbf{M} \text{ is good}] \geq 9/10$. □

B.2. Proof of Lower Bounds for AMM

Proof of Lemma 5.3. We assume that such a distribution exists with $m \leq c/\varepsilon^2$ for a small enough constant c. Let $\mathbf{A} \in \mathbb{R}^{d\times n}$ be a uniformly random orthonormal matrix ($\mathbf{A}^\mathsf{T}\mathbf{A} = I_n$) independent of the sketching matrix. Then we have

$$\Pr_{\mathbf{A},\mathbf{S}}[\|\mathbf{A}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{A} - I\|_\mathsf{F} \leq \varepsilon n] \geq 0.9,$$

as $\|A^\mathsf{T}\|_\mathsf{F} = \sqrt{n}$. Let $\mathbf{S} = U\Sigma V^\mathsf{T}$ be the singular value decomposition with $U \in \mathbb{R}^{m\times m}$, $\Sigma \in \mathbb{R}^{m\times m}$ and $V^\mathsf{T} \in \mathbb{R}^{m\times d}$. Note that if $\mathbf{S}$ is a random matrix that satisfies the AMM property, then $\mathbf{S}\cdot\mathbf{Q}$ is also a random matrix that satisfies the AMM property where $\mathbf{Q}$ is an independent uniformly random orthonormal matrix. Therefore, we can without loss of generality assume that Σ is independent of V^T and that V^T is a uniformly random orthonormal matrix. Thus, the above condition implies that

$$\Pr_{\mathbf{A},V,\Sigma}[\|\mathbf{A}^\mathsf{T}V\Sigma^2V^\mathsf{T}\mathbf{A} - I\|_\mathsf{F} \leq \varepsilon n] \geq 0.9.$$

Using the following lemma, we effectively show that the matrix $V^\mathsf{T}\mathbf{A}$ in the above statement can be replaced with $(1/\sqrt{d})\hat{\mathbf{G}}$, where $\hat{\mathbf{G}}$ is a Gaussian matrix of the same dimensions as $V^\mathsf{T}\mathbf{A}$.

Lemma B.3 (Lemma 3 of Li & Woodruff (2021))**.** *Let $\mathbf{G} \sim \mathcal{G}_{d,d}$ and $\mathbf{Z} \sim O^{d\times d}$. Suppose that $p, q \leq d$ and $\hat{\mathbf{G}}$ is the top left $p \times q$ block of $\mathbf{G}$ and $\hat{Z}$ is the top left $p \times q$ block of Z. Then $d_{KL}(\frac{1}{\sqrt{d}}\hat{\mathbf{G}}\|\hat{Z}) \leq C\frac{pq}{d}$, where C is a universal constant. By applying Pinsker's inequality, we obtain that*

$$d_{TV}(\frac{1}{\sqrt{d}}\hat{\mathbf{G}}\|\hat{Z}) \leq \sqrt{(1/2)d_{KL}(\frac{1}{\sqrt{d}}\hat{\mathbf{G}}\|\hat{Z})} \leq \sqrt{Cpq/2d}.$$

Now both the matrices $V, \mathbf{A}$ can be taken to be the first m and n columns of independent uniform random orthogonal matrices $\mathbf{V}'$ and $\mathbf{A}'$, respectively. By the properties of the Haar Measure, we obtain that $\mathbf{V}'^\mathsf{T}\mathbf{A}'$ is also a uniform random orthogonal matrix. Thus, the matrix $V^\mathsf{T}\mathbf{A}$ can be seen as the top left $m \times n$ sub-matrix of a uniformly random orthogonal matrix. If $nm \leq d/100C$, which can be assumed as $m \leq c/\varepsilon^2$ for a small enough constant α, we have from the above lemma that $d_{\mathrm{TV}}(\frac{1}{\sqrt{d}}\mathbf{G}\|V^\mathsf{T}\mathbf{A}) \leq 0.1$ which implies that

$$|\Pr_{\mathbf{G},\Sigma}[\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F} \leq \varepsilon n] - \Pr_{\mathbf{A},V,\Sigma}[\|\mathbf{A}^\mathsf{T}V\Sigma^2V^\mathsf{T}\mathbf{A} - I\|_\mathsf{F} \leq \varepsilon n]| \leq 0.1$$

and therefore

$$\Pr_{\mathbf{G},\Sigma}[\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F} \leq \varepsilon n] \geq 0.8, \tag{5}$$

where $\mathbf{G}$ is an $m \times n$ matrix of i.i.d. normal random variables. We will now show that if $m \ll 1/\varepsilon^2$, then no distribution over matrices Σ satisfies the above condition. Note that $\mathbf{G}$ and Σ are independent. We prove this by showing that a random matrix Σ satisfying the above probability statement must satisfy two properties simultaneously that cannot be satisfied unless $m \geq c/\varepsilon^2$ for a large enough constant c.

Let $\mathbf{G}_l$ denote the left half of the matrix $\mathbf{G}$ and $\mathbf{G}_r$ denote the right half of the matrix $\mathbf{G}$. We have

$$\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F}^2 \geq \frac{1}{d^2}\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2$$

which is obtained by considering the Frobenius norm of only the bottom-left block matrix of $(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I$. We first have the following lemma.

Lemma B.4. *Let M be a fixed matrix and $\mathbf{G}$ be a Gaussian matrix with t columns. Then with probability ≥ 0.9, $\|M\mathbf{G}\|_\mathsf{F}^2 \geq 0.001t\|M\|_\mathsf{F}^2$.*

Proof. Let $M = U\Sigma V^\mathsf{T}$. We have $M\mathbf{G} = U\Sigma V^\mathsf{T}\mathbf{G} = U\Sigma\mathbf{G}'$ where $\mathbf{G}'$ is a Gaussian matrix with t columns. Now, $\|M\mathbf{G}\|_\mathsf{F}^2 = \|U\Sigma\mathbf{G}'\|_\mathsf{F}^2 = \|\Sigma\mathbf{G}'\|_\mathsf{F}^2 = \sum_i\sum_j \sigma_i^2 g_{ij}^2$. By Lemma B.2, $\sum_i\sum_j \sigma_i^2 g_{ij}^2 \geq (\sum_i\sum_j \sigma_i^2)\cdot 0.001$ with probability ≥ 0.9. Now, using the equality $\sum_i\sum_j \sigma_i^2 = \sum_i t\sigma_i^2 = t\|\Sigma\|_\mathsf{F}^2 = t\|M\|_\mathsf{F}^2$, we finish the proof. ☐

Thus, conditioned on the matrix $\mathbf{G}_r^\mathsf{T}\Sigma^2$, we have that with probability ≥ 0.9,

$$\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2 \geq 0.001(n/2)\|\mathbf{G}_r^\mathsf{T}\Sigma^2\|_\mathsf{F}^2.$$

Applying the same lemma again, we have with probability ≥ 0.9, $\|\mathbf{G}_r^\mathsf{T}\Sigma^2\|_\mathsf{F}^2 \geq 0.001(n/2)\|\Sigma^2\|_\mathsf{F}^2$. Thus, overall with probability ≥ 0.8 over $\mathbf{G}$, we have for any fixed Σ that, $\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2 \geq \Omega(n^2\|\Sigma^2\|_\mathsf{F}^2)$. and therefore,

$$\Pr_{\mathbf{G},\Sigma}[\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2 \geq \Omega(n^2\|\Sigma^2\|_\mathsf{F}^2)] \geq 0.8.$$

Using a union bound with (5), we obtain that with probability ≥ 0.6, it is simultaneously true that

$$\varepsilon^2 n^2 \geq \|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F}^2 \geq \frac{1}{d^2}\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2$$

and

$$\|\mathbf{G}_r^\mathsf{T}\Sigma^2\mathbf{G}_l\|_\mathsf{F}^2 \geq \Omega(n^2\|\Sigma^2\|_\mathsf{F}^2)$$

which implies that with probability ≥ 0.6, $(1/d^2)\|\Sigma^2\|_\mathsf{F}^2 = O(\varepsilon^2)$ i.e., $(1/d^2)\sum_{i=1}^m \sigma_i^4 = O(\varepsilon^2)$. Thus, if $\mathbf{S}$ is a random matrix that satisfies the AMM property and if $\sigma_1, \ldots, \sigma_r$ are its singular values, then with probability ≥ 0.6,

$$\sum_i \sigma_i^4 \leq C_1 d^2 \varepsilon^2 \tag{6}$$

for a universal constant C_1.

We now obtain a different probability statement about the singular values of the sketching matrix $\mathbf{S}$ by considering the sum of squares of the diagonal entries of the matrix $(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I = (1/d)\sum_{i=1}^m \sigma_i^2 \mathbf{g}_i\mathbf{g}_i^\mathsf{T} - I$ where $\mathbf{g}_i$ are n dimensional Gaussian vectors. Note that $((1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I)_{jj} = (1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_{ij}^2 - 1$. Fix the matrix Σ. Clearly,

$$\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F}^2 \geq \sum_{j=1}^n((1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_{ij}^2 - 1)^2.$$

If $\sum_{i=1}^m \sigma_i^2 \leq d/100$, we have that with probability at least 0.9, $(1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_{ij}^2 \leq (10/d)\mathbb{E}[\sum_{i=1}^m \sigma_i^2\mathbf{g}_{ij}^2] \leq (10/d)\cdot(d/100)$ which implies that $((1/d)\sum_{i=1}^m \sigma_i^2\mathbf{g}_{ij}^2 - 1)^2 \geq 1/4$ with probability ≥ 0.9. Let $j \in [n]$ be *large* if the previous event holds. By a Chernoff bound, with probability ≥ 0.9, there are $\geq n/C_2$ *large* values $j \in [n]$ for a large enough absolute constant C_2. Thus, $\sum_{i=1}^m \sigma_i^2 \leq d/100$ implies that with probability ≥ 0.9, $\|(1/d)G^\mathsf{T}\Sigma^2 G - I\|_\mathsf{F}^2 \geq (n/C_2)(1/4) = n/4C_2 \geq \varepsilon^2 n^2$ as we assumed that $\varepsilon \leq c/\sqrt{n}$ for a small enough constant c. Now, if $\Pr_\Sigma[\sum_{i=1}^m \sigma_i^2 \leq d/100] > 0.3$, then by the above property for a fixed Σ, $\Pr_{\Sigma,G}[\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F}^2 \geq \varepsilon^2 n^2] > 0.2$ which implies that

$$\Pr_{\Sigma,\mathbf{G}}[\|(1/d)\mathbf{G}^\mathsf{T}\Sigma^2\mathbf{G} - I\|_\mathsf{F}^2 \leq \varepsilon^2 n^2] < 0.8$$

which is a contradiction to (5). Thus, $\Pr_\Sigma[\|\Sigma\|_\mathsf{F}^2 \leq d/100] < 0.3$ which implies

$$\Pr_\Sigma[\sum_{i=1}^m \sigma_i^2 \geq d/100] \geq 0.7. \tag{7}$$

By a union bound on (6) and (7), with probability ≥ 0.3, it simultaneously holds that

$$\sum_{i=1}^m \sigma_i^4 \leq C_1 d^2\varepsilon^2 \text{ and } \sum_{i=1}^m \sigma_i^2 \geq d/100.$$

Now,

$$d^2/100^2 \leq \left(\sum_{i=1}^m \sigma_i^2\right)^2 \leq m\sum_{i=1}^m \sigma_i^4 \leq C_1 m d^2\varepsilon^2.$$

Here we used the Cauchy-Schwarz inequality which finally implies that $m = \Omega(1/\varepsilon^2)$. Thus, any oblivious distribution that gives AMM with $\varepsilon < c/\sqrt{n}$ for a small enough constant c must have $\Omega(1/\varepsilon^2)$ rows. □

Before proving Theorem 5.4, we first prove the following lemma that shows CountSketch preserves the Frobenius norm of a matrix.

Lemma B.5 (CountSketch Preserves Frobenius Norms)**.** *If $\mathbf{S}$ is a CountSketch matrix with $m \geq 200/\varepsilon^2$, then for any arbitrary matrix A, with probability $\geq 9/10$,*

$$\|\mathbf{S}A\|_{\mathsf{F}}^2 = (1 \pm \varepsilon)\|A\|_{\mathsf{F}}^2.$$

Proof. For any vector x, we have $\mathbb{E}[(\|\mathbf{S}x\|_2^2 - \|x\|_2^2)^2] \leq (2/m)\|x\|_2^4$ if $\mathbf{S}$ is a CountSketch matrix with m rows. Now,

$$\mathbb{E}[|\|\mathbf{S}x\|_2^2 - \|x\|_2^2|]^2 \leq \mathbb{E}[(\|\mathbf{S}x\|_2^2 - \|x\|_2^2)^2] \leq (2/m)\|x\|_2^4$$

which implies that $\mathbb{E}[|\|\mathbf{S}x\|_2^2 - \|x\|_2^2|] \leq \sqrt{2/m}\|x\|_2^2$. For any arbitrary matrix A, we have $|\|\mathbf{S}A\|_2^2 - \|A\|_2^2| = |\sum_i \|\mathbf{S}A_{*i}\|_{\mathsf{F}}^2 - \|A_{*i}\|_{\mathsf{F}}^2| \leq \sum_i |\|\mathbf{S}A_{*i}\|_2^2 - \|A_{*i}\|_2^2|$. Thus, $\mathbb{E}[|\|\mathbf{S}A\|_{\mathsf{F}}^2 - \|A\|_{\mathsf{F}}^2|] \leq \sum_i \mathbb{E}[|\|\mathbf{S}A_{*i}\|_2^2 - \|A_{*i}\|_2^2|] \leq \sqrt{2/m}\sum_i \|A_{*i}\|_2^2 = \sqrt{2/m}\|A\|_{\mathsf{F}}^2$. For $m \geq 200/\varepsilon^2$, we have $\mathbb{E}[|\|\mathbf{S}A\|_{\mathsf{F}}^2 - \|A\|_{\mathsf{F}}^2|] \leq (\varepsilon/10)\|A\|_{\mathsf{F}}^2$. By Markov's inequality, with probability $\geq 9/10$, $\|\mathbf{S}A\|_{\mathsf{F}}^2 = (1 \pm \varepsilon)\|A\|_{\mathsf{F}}^2$. □

Proof of Theorem 5.4. Given n and $\varepsilon \leq c/\sqrt{n}$ for a small enough constant, assume that for $d = C_1/\varepsilon^2$ for a large enough universal constant C_1, there is a random matrix $\mathbf{S}$ with $r < C_2/\varepsilon^2$ rows such that for any fixed matrix $A \in \mathbb{R}^{d \times n}$, with probability $\geq 99/100$,

$$\|A^{\mathsf{T}}\mathbf{S}^{\mathsf{T}}\mathbf{S}A - A^{\mathsf{T}}A\|_{\mathsf{F}} \leq (\varepsilon/3)\|A^{\mathsf{T}}\|_{\mathsf{F}}\|A\|_{\mathsf{F}}.$$

Now consider an arbitrary matrix $B \in \mathbb{R}^{d' \times n}$ for $d' \geq Cn/\varepsilon^2$ for which the previous lemma applies. Let $\mathbf{S}_1$ be a CountSketch matrix with K/ε^2 rows for a large enough K. With probability $\geq 95/100$, we simultaneously have (i) $\|B^{\mathsf{T}}\mathbf{S}_1^{\mathsf{T}}\mathbf{S}_1 B - B^{\mathsf{T}}B\|_{\mathsf{F}} \leq (\varepsilon/3)\|B^{\mathsf{T}}\|_{\mathsf{F}}\|B\|_{\mathsf{F}} = (\varepsilon/3)\|B\|_{\mathsf{F}}^2$, and (ii) $\|\mathbf{S}_1 B\|_{\mathsf{F}} = (1 \pm \varepsilon/3)\|B\|_{\mathsf{F}}$. By picking C_1 large enough, we have that $C_1 \geq K$. Thus, by our assumption, the random matrix $\mathbf{S}$ gives the AMM property for the matrix $\mathbf{S}_1 A$. Conditioning on the above events, we have with probability $\geq 99/100$ that

$$\begin{aligned}
&\|B^{\mathsf{T}}\mathbf{S}_1^{\mathsf{T}}\mathbf{S}^{\mathsf{T}}\mathbf{S}\mathbf{S}_1 B - B^{\mathsf{T}}\mathbf{S}_1^{\mathsf{T}}\mathbf{S}_1 B\|_{\mathsf{F}} \\
&\leq (\varepsilon/3)\|B^{\mathsf{T}}\mathbf{S}_1^{\mathsf{T}}\|_{\mathsf{F}}\|\mathbf{S}_1 B\|_{\mathsf{F}} \\
&\leq (\varepsilon/3)(1+\varepsilon/3)^2\|B\|_{\mathsf{F}}^2 \leq (\varepsilon/2)\|B\|_{\mathsf{F}}^2.
\end{aligned}$$

By the triangle inequality, we obtain $\|B^{\mathsf{T}}\mathbf{S}_1^{\mathsf{T}}\mathbf{S}^{\mathsf{T}}\mathbf{S}\mathbf{S}_1 B - B^{\mathsf{T}}B\|_{\mathsf{F}} \leq (\varepsilon/3 + \varepsilon/2)\|B\|_{\mathsf{F}}^2 \leq \varepsilon\|B\|_{\mathsf{F}}^2$. Thus, by a union bound, with probability ≥ 0.9, the random matrix $\mathbf{S} \cdot \mathbf{S}_1$ satisfies that for any fixed matrix B with $\geq Cn/\varepsilon^2$ rows, with probability ≥ 0.9,

$$\|B^{\mathsf{T}}(\mathbf{S} \cdot \mathbf{S}_1)^{\mathsf{T}}(\mathbf{S} \cdot \mathbf{S}_1)B - B^{\mathsf{T}}B\|_{\mathsf{F}} \leq \varepsilon\|B\|_{\mathsf{F}}^2$$

implying that even for a matrix with at least Cn/ε^2 rows, there is an oblivious sketching distribution with $r < C_2/\varepsilon^2$ rows which gives an AMM guarantee. This contradicts the previous lemma and hence our assumption that there is a small sketching matrix for matrices with $d = C_1/\varepsilon^2$ rows is false. Thus, we have that for any $d \geq C/\varepsilon^2$ for a large enough constant C, there is no sketching distribution with $< C_1/\varepsilon^2$ rows that gives the ε AMM guarantee for matrices with $\geq d$ rows. □

C. Missing Proofs from Section 6

C.1. Lower Bounds for Ridge Regression

Note that multiplying a column of M with -1 does not change the optimum value. Therefore, without loss of generality, we can assume that the columns of M are either $[1\,1]^{\mathsf{T}}$ or $[1\,-1]^{\mathsf{T}}$. Thus, we have that $d - N$ columns of M are equal to $[1\,1]^{\mathsf{T}}$ and N columns of the matrix M are equal to $[1\,-1]^{\mathsf{T}}$. Let $I \subseteq [n]$, $|I| = N$ be the set of columns of M that are equal to

$[1\ -1]^\mathsf{T}$. Now, for any vector x,

$$\begin{aligned}
&\|Mx - \begin{bmatrix}1\\-1\end{bmatrix}\|_2^2 + \lambda\|x\|_2^2 \\
&= (\sum_{i\in I} x_i + \sum_{i\notin I} x_i - 1)^2 + (-\sum_{i\in I} x_i + \sum_{i\notin I} x_i + 1)^2 \\
&\quad + \lambda\sum_{i\in I} x_i^2 + \lambda\sum_{i\notin I} x_i^2 \\
&= 2(\sum_{i\notin I} x_i)^2 + 2(\sum_{i\in I} x_i - 1)^2 + \lambda\sum_{i\in I} x_i^2 + \lambda\sum_{i\notin I} x_i^2.
\end{aligned}$$

Clearly, to minimize the expression, we need to set $x_i = 0$ for $i \notin I$. We also have that for a fixed value of $\sum_{i\in I} x_i$, the expression $\sum_{i\in I} x_i^2$ is minimized when $x_i = x_{i'}$ for all $i, i' \in I$. Let $x_i = \alpha$ for all $i \in I$. Then,

$$\|Mx - \begin{bmatrix}1\\-1\end{bmatrix}\|_2^2 + \lambda\|x\|_2^2 = 2(N\alpha - 1)^2 + \lambda N\alpha^2.$$

To minimize the expression, we set $\alpha = 2/(2N + \lambda)$ and obtain that

$$\min_x \|Mx - \begin{bmatrix}1\\-1\end{bmatrix}\|_2^2 + \lambda\|x\|_2^2 = \frac{2\lambda}{\lambda + 2N}.$$

D. Communication Lower Bounds for other Regularized Problems

We show communication lower bounds for computing $1+\varepsilon$ approximations to optimum values of Lasso and the square-root Lasso problems by reducing the DISJOINTNESS problem to computing these optimum values.

In an instance of DISJOINTNESS problem Alice receives a set $A \subseteq [n]$ and Bob receives a set $B \subseteq [n]$. They send bits to each other to communicate based on a pre-determined to protocol to find if $A \cap B = \emptyset$ or not.

Theorem D.1 (Hardness of DISJOINTNESS (Håstad & Wigderson, 2007))**.** *Every public-coin randomized protocol for* DISJOINTNESS *that has two-sided error at most a constant $\varepsilon \in (0, 1/2)$ uses $\Omega(n)$ communication in the worst case (over inputs and coin flips).*

Theorem D.2 (Communication Complexity of Lasso)**.** *Let $0 < \lambda < 1$ be the Lasso parameter. If Alice has the $n \times d$ matrix M_1 and Bob has the $n \times d$ matrix M_2, then to determine a $1 + c$ approximation, for a small enough constant c, to the optimal value of*

$$\min_z \|(M_1 + M_2)z - b\|_2^2 + \lambda\|z\|_1$$

requires $\Omega(d)$ bits of communication between Alice and Bob.

Proof. We prove the theorem by reducing the DISJOINTNESS problem to an instance of the Lasso problem. Let $A, B \subseteq [d]$ denote the sets received by Alice and Bob, respectively. Let $x^{(1)}, x^{(2)} \in \{0,1\}^d$ be the vectors that denote A and B respectively. We have $x_i^{(1)} = 1$ if and only if $i \in A$ and similarly $x_i^{(2)} = 1$ if and only if $i \in B$. Let M denote a $2 \times d$ matrix with two rows such that $M_{1*} = x^{(1)}$ and $M_{2*} = x^{(2)}$. We consider the Lasso problem:

$$\min_z \|Mz - \begin{bmatrix}1\\1\end{bmatrix}\|_2^2 + \lambda\|z\|_1$$

for some $0 < \lambda < 1$.

Consider the following two cases:

1. $A \cap B \neq \emptyset$: Let $i \in A \cap B$. We have $x_i^{(1)} = x_i^{(2)} = 1$ and therefore $M_{1i} = M_{2i} = 1$. Let $z = e_i$. We have $Me_i = [1 \quad 1]^\mathsf{T}$. Therefore

$$\|Me_i - \begin{bmatrix}1\\1\end{bmatrix}\|_2^2 + \lambda\|e_i\|_1 = 0 + \lambda = \lambda.$$

Therefore $\min_z \|Mz - 1_2\|_2^2 + \lambda\|z\|_1 \leq \lambda$.

2. Let $A \cap B = \emptyset$. Let z be an arbitrary vector. We have $M_{1*}z = \sum_{i \in A} z_i$ and $M_{2*}z = \sum_{i \in B} z_i$. Thus,

$$\|Mz - 1_2\|_2^2 + \lambda\|z\|_1 = (\sum_{i \in A} z_i - 1)^2 + (\sum_{i \in B} z_i - 1)^2 + \lambda(\sum_{i \in A} |z_i| + \sum_{i \in B} |z_i|),$$

We can see that the optimum value can be attained by setting $z_i = \alpha$ for some $i \in A$ and $z_j = \beta$ for some $j \in B$ and setting the rest of the coordinates to 0. We now consider

$$\min_{\alpha} (\alpha - 1)^2 + \lambda\alpha.$$

The optimal solution for the above is attained at α satisfying $2(\alpha - 1) + \lambda = 0$ which gives $\alpha = 1 - \lambda/2$. Thus the optimum value for

$$\min_{z} \|Mz - 1_2\|_2^2 + \lambda\|z\|_1$$

when $A \cap B = \emptyset$ is given by

$$2 \cdot \frac{\lambda^2}{4} + \lambda(2 - \lambda) = 2\lambda - \lambda^2/2 \geq 3\lambda/2.$$

The optimal Lasso values differ by a factor of at least $3/2$. Thus a protocol that can compute $1 + c$ approximation to the optimal value of ridge regression can be used to solve the DISJOINTNESS problem which gives an $\Omega(d)$ bits lower bound for approximating Lasso. □

We similarly have the following communication lower bounds for Square-root Lasso.

Theorem D.3 (Hardness of Sketching Square-Root Lasso)**.** *Let $0 < \lambda < 2\sqrt{2}/3$ be the Lasso parameter. If Alice has the $n \times d$ matrix M_1 and Bob has the $n \times d$ matrix M_2, then to determine a $1 + c$ approximation, for a small enough constant c, to the optimal value of*

$$\min_{z} \|(M_1 + M_2)z - b\|_2 + \lambda\|z\|_1,$$

requires $\Omega(d)$ bits of communication between Alice and Bob.

Proof. We use the same notation as in the proof of the above theorem. In the case of $A \cap B \neq \emptyset$, we again have that the optimum value is at most λ. In the other case, where $A \cap B = \emptyset$, we again have that the optimum value can be attained by setting $z_i = \alpha = z_j$ for some $i \in A$ and $j \in B$. Thus we have the following optimization problem

$$\min_{\alpha} \sqrt{2}|\alpha - 1| + 2\lambda|\alpha| = \min(\sqrt{2}, 2\lambda).$$

Thus, the optimum value of the square-root Lasso problem is at most λ in the case $A \cap B \neq \emptyset$ and is at least $\min(\sqrt{2}, 2\lambda)$ in the case of $A \cap B = \emptyset$. If $\lambda < 2\sqrt{2}/3$, then $\min(\sqrt{2}, 2\lambda) \geq (3\lambda/2)$. Thus, if Alice and Bob can obtain an $11/10$ approximation for the Square-Root Lasso instance $(M, 1_2, \lambda)$ for some $\lambda < 2\sqrt{2}/3$, they can solve the DISJOINTNESS instance which implies an $\Omega(d)$ bit lower bound for computing a $1 + c$ approximation for Square-Root Lasso. □

E. An Experiment

We run our algorithm on a ridge regression instance with a 6000×70000 matrix A whose entries are independent Gaussian random variables. We set λ such that $\sigma^2/\lambda \approx 1$. Naïvely computing $x^* = A^\mathsf{T}(AA^\mathsf{T} + \lambda I)^{-1}b$ takes $t_{\text{naive}} = 71$ seconds on our machine. We use OSNAP with sparsity $s = 8$ and vary the number r of rows and observe the running times and quality of the solution that is obtained by our algorithm.

Our experiments show the general trends we expect. Increasing the number of rows in the sketching matrix results in a solution $\tilde{x}$ that has lower cost and also is closer to the optimum solution x^*. We also see that the running time of the algorithm is nearly linear in the sketch size r, implying that the time required to apply the sketch is negligible for this instance. At sketch size $r = 30000$, that is less than $d/2$, we see that the algorithm runs nearly 40% faster than the naïve algorithm while computing a solution that has a cost within 5% of the optimum. For larger values of d, we expect to obtain a greater speedup as compared to naïvely computing x^*.

Notice that we do not compare with the algorithm of Chowdhury et al. (2018) as for one iteration, our algorithm is exactly the same as theirs. Our theorems show that the sketch can be smaller and sparser than what is shown in their work to compute $1+\varepsilon$ approximate solutions, giving the first proof of correctness about the quality of the solution at smaller sketch sizes.

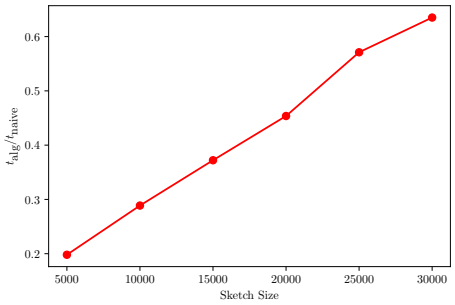

Figure 1. $t_{\text{alg}}/t_{\text{naive}}$ vs # of rows of OSNAP

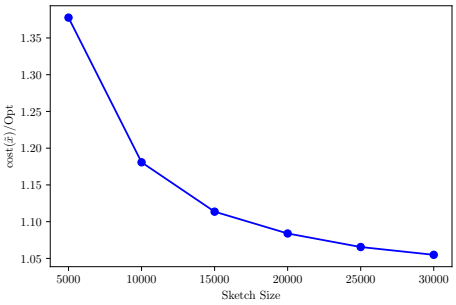

Figure 2. $\text{cost}(\tilde{x})/\text{Opt}$ vs # of rows of OSNAP

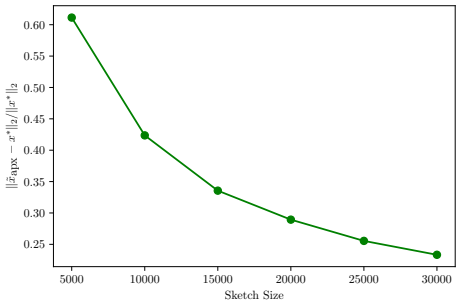

Figure 3. $\|\tilde{x}-x^*\|_2/\|x^*\|_2$ vs # of rows of OSNAP