# Optimal Deterministic Coresets for Ridge Regression

**Praneeth Kacham**
Carnegie Mellon University

**David P. Woodruff**
Carnegie Mellon University

## Abstract

We consider the ridge regression problem, for which we are given an $n \times d$ matrix $A$ of examples and a corresponding $n \times d'$ matrix $B$ of labels, as well as a ridge parameter $\lambda \geq 0$, and would like to output an $X' \in \mathbb{R}^{d \times d'}$ for which

$$\|AX' - B\|_F^2 + \lambda \|X'\|_F^2 \leq (1+\epsilon)\text{OPT},$$

where $\text{OPT} = \min_{Y \in \mathbb{R}^{d \times d'}} \|AY - B\|_F^2 + \lambda \|Y\|_F^2$. In the special case of $\lambda = 0$, this is ordinary multi-response linear regression. Our focus is on deterministically constructing coresets for this problem. Here the goal is to select and re-weight a small subset of rows of $A$ and corresponding labels of $B$, denoted by $SA$ and $SB$, so that if $X'$ is the minimizer to $\min_{X'} \|SAX' - SB\|_F^2 + \lambda \|X'\|_F^2$, then $\|AX' - B\|_F^2 + \lambda \|X'\|_F^2 \leq (1+\epsilon)\text{OPT}$.

We show how to efficiently (in $\text{poly}(n, d, 1/\epsilon)$ time) and deterministically select $O(\texttt{sd}_\lambda/\epsilon)$ rows of $A$ and $B$ to achieve this property, and prove a matching lower bound, showing that it is necessary to select $\Omega(\texttt{sd}_\lambda/\epsilon)$ rows no matter what the weights are, for any $1 < 1/\epsilon \leq \texttt{sd}_\lambda$. Here $\texttt{sd}_\lambda$ is the statistical dimension of the input, and we assume $d' = O(\texttt{sd}_\lambda) \leq d$. In the case of ordinary regression, this gives a deterministic algorithm achieving $O(d/\epsilon)$ rows and a matching lower bound for any $1 \leq 1/\epsilon \leq d$; for $1/\epsilon > d$ we show $\Theta(d^2)$ rows are sufficient. Finally we show our new coresets are mergeable, giving a deterministic protocol for ridge regression with $O(\texttt{sd}_\lambda/\epsilon)$ words of communication per server in a distributed setting, in the important case when the rows of $A$ and $B$ have a constant number of non-zero entries and there are a constant number of servers. Prior to our work the best deterministic protocols in this setting required $\Omega(\min(\texttt{sd}_\lambda^2, \texttt{sd}_\lambda/\epsilon^2))$ communication.

## 1 Introduction

Linear least squares regression is one of the most popular tools for fitting a linear hypothesis to a given data set and ridge regression is an important regularized variant. When the number $n$ of data points is very large, an intriguing question is whether there exists a small weighted subset of the data points which represents the entire data well for ridge regression. These subsets are often called *coresets*.

Let $A$ be an $n \times d$ input matrix and $B$ an $n \times d'$ matrix of labels corresponding to the data points in $A$. Here each label is a $d'$-dimensional vector. Let $a_i \in \mathbb{R}^d$ denote the $i$-th row of a $A$ written as a column. In the multiple-response least squares regression problem, the goal is to find a matrix $X$ such that $\|AX - B\|_F^2$ is minimized, where for a matrix $C$, the squared Frobenius norm $\|C\|_F^2$ is the sum of squares of its entries. In the ridge regression problem, we additionally add an $\ell_2$-regularizer to the cost and now the goal is to find a matrix $X$ which minimizes $\|AX - B\|_F^2 + \lambda \|X\|_F^2$, where $\lambda > 0$ is the regularization parameter.

We call a subset $S \subseteq [n]$, along with corresponding weights $w_i \geq 0$ for $i \in S$, an $\epsilon-$coreset if the solution to the ridge regression problem

$$\tilde{X}_{S,w} = \underset{X}{\operatorname{argmin}} \sum_{i \in S} w_i \|a_i^T X - b_i\|_2^2 + \lambda \|X\|_F^2$$

is a $(1+\epsilon)$-approximate solution to the ridge regression problem $\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2 = \min_X \sum_i \|a_i^T X - b_i\|_2^2 + \lambda \|X\|_F^2$ i.e.,

$$\begin{aligned} &\|A\tilde{X}_{S,w} - B\|_F^2 + \lambda \|\tilde{X}_{S,w}\|_F^2 \\ &\leq (1+\epsilon)\left(\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2\right). \end{aligned}$$

Ideally, we would like to have the size $|S|$ of $S$ be independent of $n$ and depend linearly on the dimension

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

of the data $d$ and sub-linear in $1/\epsilon$. In the case of ridge regression, it is often desirable to have bounds in terms of the *statistical dimension* $\mathtt{sd}_\lambda$ of the input (defined below), which is always at most $d$ and often significantly smaller than $d$.

Obtaining small subsets which accurately represent the entire data set is crucial for data interpretation and for efficient communication protocols. Note that unlike other solutions, such as directly computing the covariance matrix, coresets preserve the sparsity of the data. Indeed, if the rows of $A$ and $B$ are sparse, then the selected rows in the coreset are also sparse. As we will see, small coresets are extremely useful in giving efficient communication protocols to solve problems in a distributed setting.

In this work we focus on *deterministic* algorithms, i.e., algorithms with zero error probability. Since coresets are often composed multiple times in distributed protocols, this is desirable so that the error probability does not compound. Moreover, a deterministic coreset allows one to generate additional rows of $A$ and $B$ as a function of the data and coresets we have already computed. This allows one to *adaptively* generate data. Note that in general this is not possible if the coreset is randomized. Indeed, if the input to a coreset is allowed to depend on the randomness used to build the coreset, there are no guarantees.

### 1.1 Previous Work

There is a vast body of work on least squares and ridge regression, and we only touch upon the works most relevant to ours here and refer the reader to the surveys (Mahoney, 2011; Woodruff, 2014) and references therein. There is a long line of work on randomized sampling algorithms for speeding up least squares regression, see, e.g., (Drineas et al., 2006a,b,c, 2011). Since our focus here is on deterministic algorithms, these are not directly useful for us. In the unregularized case, a direct technique that we can apply is the deterministic spectral sparsification result of Batson, Spielman, and Srivastava (BSS) (Batson et al., 2012). There are also several followup works (Allen Zhu et al., 2015; Lee and Sun, 2018; Chen and Price, 2019), but they give randomized rather than deterministic algorithms.

Assume $d' = O(\mathtt{sd}_\lambda) \leq d$. The issue with directly using the BSS algorithm for ordinary least squares regression is that naïvely one would need a so-called subspace embedding of the column span of $C = [A, B]$, the matrix with the columns of $B$ adjoined to those of $A$. Consequently, this would result in a coreset $S$ containing $O(d/\epsilon^2)$ rows, which is larger than the $O(d/\epsilon)$ that we desire. We instead achieve $O(d/\epsilon)$ rows by combining the deterministic guarantees needed for regression in (Avron et al., 2017) with a deterministic row selection algorithm achieving approximate matrix product in (Cohen et al., 2016). Using this property, we can then bootstrap from it to in turn obtain a coreset of size $O(\mathtt{sd}_\lambda/\epsilon)$. Directly applying techniques in (Avron et al., 2017) would instead result in a coreset containing $O(\mathtt{sd}_\lambda/\epsilon^2)$ rows.

Previous work (Maalouf et al., 2019) has also observed that one can preserve the covariance matrix $C^T C$ exactly by a coreset of $O(d^2)$ rows by using Caratheodory's theorem, which can be implemented in deterministic polynomial time. However, it was not known if there is a matching lower bound in the case of least squares regression. There are strong lower bounds for cut and spectral sparsifiers (Andoni et al., 2016; Carlson et al., 2017); however, they fail to apply to the case of regression when there is a specific $B$ matrix given.

There is also a body of work on distributed regression, for which each of the rows of $C = [A, B]$ reside on a single server. We refer the reader to the recent work (Vempala et al., 2019) and references therein. As shown in (Vempala et al., 2019), for ordinary least squares regression, $\Theta(d^2)$ words of communication per server is necessary and sufficient to solve the problem up to any relative error accuracy. The protocol is simple - each server computes its local covariance matrix and sends it to the coordinator, who can then solve the least squares problem exactly. While (Vempala et al., 2019) proves this is optimal, even to obtain a constant factor approximation, it need not be optimal if each row of $C$ only has $O(1)$ non-zero entries. In this case one could hope to do better than $d^2$ communication by transmitting a small number of rows. We note that by Caratheordory's theorem, one can still transmit $O(d^2)$ rows or $O(\mathtt{sd}_\lambda^2)$ rows for the regularized version, assuming $d' \leq \mathtt{sd}_\lambda$, but the hope is to do even better. Alternatively, one can transmit a subspace embedding using $O(d/\epsilon^2)$ rows, or $O(\mathtt{sd}_\lambda/\epsilon^2)$ rows for the regularized version, but these are not linear in $1/\epsilon$. Alternatively, one could use one of many randomized algorithms (see, e.g, (Woodruff, 2014)) to obtain $O(d/\epsilon)$ or $O(\mathtt{sd}_\lambda/\epsilon)$ communication by using an oblivious sketch; however, these cannot be made deterministic. Thus, an interesting question arises if there is a deterministic protocol achieving better communication. To the best of our knowledge, such work has not considered the sparse case, i.e., when each row of $A$ and corresponding row of $B$ have at most $O(1)$ non-zero entries.

## 1.2 Our Contributions

Given matrices $A \in \mathbb{R}^{n \times d}, B \in \mathbb{R}^{n \times d'}$ and parameter $\lambda$, we give a *deterministic* algorithm to find an $\epsilon$-coreset $S$ of size $O((\mathtt{sd}_\lambda(A)+d')/\epsilon)$ and corresponding weights. We do this by using Corollary 1 from Cohen et al. (2016) on suitably defined matrices and show that the matrix $S$ thus obtained defines an $\epsilon$-coreset for the ridge regression problem. This immediately gives that, with parameter $\lambda = 0$, there is an $\epsilon$-coreset of size $O((\mathtt{rank}(A) + d')/\epsilon)$.

**Theorem 1.1.** *Given matrices $A \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{n \times d'}$ and $\lambda \geq 0$, there exists a matrix $S$ which selects and scales $O((\mathtt{sd}_\lambda + d')/\epsilon)$ rows of $A$ such that solution to the ridge regression problem*

$$\min_X \|SAX - SB\|_2^2 + \lambda \|X\|_F^2$$

*is a $(1+\epsilon)$ approximate solution to the ridge regression problem*

$$\min_x \|AX - B\|_F^2 + \lambda \|X\|_F^2.$$

Using $\epsilon$-coresets, we give an efficient communication protocol for computing a $1+\epsilon$ approximate solution to multi-response ridge regression in a distributed setting with communication complexity of $O(s \cdot t \cdot (\min(s \cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A)) + d')/\epsilon)$ words where $s$ is the number of servers and $t$ is the maximum number of non-zero elements in a row of $[A, B]$. In the case of $t \ll d$, this protocol is much more efficient than $d^2$ words which corresponds to naïvely sending the matrices $A_i^T A_i$ to the central server.

**Theorem 1.2.** *If rows of matrix $A \in \mathbb{R}^{n \times d}$ are partitioned among $s$ servers and corresponding rows of $B \in \mathbb{R}^{n \times d'}$ are partitioned too, then there is a deterministic communication protocol using*

$$O(\frac{s \cdot t \cdot (\min(s \cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A)) + d')}{\epsilon}) \text{ words,}$$

*where $t$ is the maximum number of non-zero entries in a row of $[A, B]$.*

We finally show that our bounds on the coreset size are tight in the case of Multiple Ridge Regression for a certain setting of $\lambda$.

**Theorem 1.3.** *For all $\epsilon$ such that $1 \leq 1/100\epsilon \leq d$ and $\lambda \leq 1/4\epsilon$, there exist matrices $A, B \in \mathbb{R}^{d/100\epsilon \times d}$ for which any matrix $S$ that selects and rescales $k$ rows of $A$ and $B$ such that the solution to*

$$\min_X \|SAX - SB\|_F^2 + \lambda \|X\|_F^2$$

*is a $(1+\epsilon)$-approximation to*

$$\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2$$

*has $k = \Omega(\mathtt{sd}_\lambda(A)/\epsilon)$ rows.*

## 1.3 Notation

$A^+$ denotes the Moore-Penrose pseudo-inverse of a matrix $A$. For a matrix $A$ and a vector $x$, $\Pi_A(x)$ denotes the projection of $x$ onto the column span of $A$ given by $\Pi_A(x) = AA^+x$. $[n]$ denotes the set $\{1, 2, 3, \ldots, n\}$. $\mathtt{range}$(A) denotes the subspace spanned by the columns of the matrix $A$. We start by defining the case when $X$ just has a single column, in which case we denote $X$ by $x$.

A matrix $S$ is an $\epsilon$-subspace embedding for $\mathtt{range}(A)$, if for all $x$,

$$(1-\epsilon)\|Ax\|_2^2 \leq \|SAx\|_2^2 \leq (1+\epsilon)\|Ax\|_2^2.$$

A typical ridge regression problem is given by inputs $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$ and $\lambda \geq 0$. Let $a_i \in \mathbb{R}^n$ be the vector corresponding to the $i$-th row of matrix $A$ and $b_i \in \mathbb{R}$ be the $i$-th component of vector $b$. Let $x^*$ denote the optimum solution for ridge regression and define $\mathrm{OPT} = \|Ax^* - b\|_2^2 + \lambda\|x^*\|_2^2$. A set $S \subseteq [n]$ along with weights $w_i \geq 0$ for $i \in S$ defines the weighted ridge regression problem

$$\min_x \sum_{i \in S} w_i (a_i^T x - b_i)^2 + \lambda \|x\|_2^2.$$

Let $\tilde{x}_{S,w}$ be the optimal solution for the ridge regression problem defined by $S, w$. We say $(S, w)$ is an $\epsilon$-coreset if

$$\|A\tilde{x}_{S,w} - b\|_2^2 + \lambda\|\tilde{x}_{S,w}\|_2^2 \leq (1+\epsilon)\mathrm{OPT}.$$

For notational convenience, we define a selecting and scaling matrix $S$ corresponding to set $S \subseteq n$ and $w$, such that

$$\|SAx - Sb\|_2^2 = \sum_{i \in S} w_i (a_i^T x - b_i)^2.$$

By a selecting matrix, we mean that each row of $S$ has exactly one non-zero entry.

## 1.4 Preliminaries

### 1.4.1 Singular Value Decomposition

**Definition 1.4** (Singular Value Decomposition)**.** *Any matrix $A \in \mathbb{R}^{n \times d}$ can be written as $U \Sigma V^T$, where both $U$ and $V$ are orthonormal matrices and $\Sigma$ is a diagonal matrix with non-zero entries in a non-increasing order. We call the entries of $\Sigma$ the singular values and label them by $\sigma_1, \sigma_2 \ldots, \sigma_d$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$.*

### 1.4.2 Statistical Dimension

**Definition 1.5** (Statistical Dimension)**.** *For a matrix $A \in \mathbb{R}^{n \times d}$ with non-zero singular values $\sigma_1, \sigma_2, \ldots, \sigma_d$,* *the statistical dimension with respect to $\lambda \geq 0$, $\mathtt{sd}_\lambda(A)$, is defined to be*

$$\mathtt{sd}_\lambda(A) = \sum_{i=1}^{rank(A)} \frac{1}{1+\frac{\lambda}{\sigma_i^2}} \tag{1.1}$$

Wherever $A$ is apparent, we use the notation $\mathtt{sd}_\lambda$ for $\mathtt{sd}_\lambda(A)$. Note that $\mathtt{sd}_\lambda(A) \leq \mathtt{rank}(A) \leq d$. This definition of statistical dimension captures our intuitive notion that as $\lambda$ increases, the importance of the data decreases. Furthermore, if $\lambda \geq \sigma_1^2/\epsilon$, $0^d$ is a $1+\epsilon$ approximate solution for any ridge regression problem with data matrix $A$ and regularization parameter $\lambda$ (See Lemma 14 of (Avron et al., 2017) for a proof). Proofs of the following lemmas can be found in the supplementary material.

**Lemma 1.6.** *If $\hat{A}$ is a matrix with orthonormal columns such that $\mathtt{range}(\hat{A}) = \mathtt{range}(\begin{bmatrix} A \\ \sqrt{\lambda}I \end{bmatrix})$ and if $U_1$ comprises the first $n$ rows of $\hat{A}$, then $\|U_1\|_F^2 = \mathtt{sd}_\lambda(A)$ and $\|U_1\|_2^2 = 1/(1+\lambda/\sigma_1^2) \leq 1$.*

**Lemma 1.7.** *If $A'$ is the sub-matrix of $A$ formed by taking rows of $A$, then $\mathtt{sd}_\lambda(A') \leq \mathtt{sd}_\lambda(A)$.*

**Lemma 1.8.** *For any $r \geq 1$, $\mathtt{sd}_{\lambda/r}(A) \leq \min(r \cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A))$.*

### 1.4.3 Spectral Sparsification

**Theorem 1.9.** *(BSS Algorithm (Batson et al., 2012)) Given $n$ vectors $v_1, v_2, \ldots, v_n \in \mathbb{R}^d$, there exists a subset $S \subseteq [n]$ of size $O(d/\epsilon^2)$ with corresponding weights $w_i \geq 0$ for $i \in S$ such that*

$$(1-\epsilon)\sum_{i=1}^n v_i v_i^T \preceq \sum_{i\in S} w_i v_i v_i^T \preceq (1+\epsilon)\sum_{i=1}^n v_i v_i^T$$

*and there is a deterministic polynomial time algorithm to find this subset along with the corresponding weights.*

## 2 Upper Bounds for Linear Regression

In this section, we show that there exists an $\epsilon$-coreset of size $O(d/\epsilon)$ for "linear regression" in the single response case (when $X$ has one column) and $O((d+d')/\epsilon)$ in the multiple response case and show that the BSS algorithm can be used to find this coreset *deterministically.*

### 2.1 Single Response Linear Regression

**Lemma 2.1** (Lemma 1 of (Cohen et al., 2016))**.** *If $S$ is an $\epsilon$-subspace embedding for $\mathtt{colspan}(A, B)$,*

$$\|A^T S^T S B - A^T B\|_2 \leq \epsilon \|A\|_2 \|B\|_2.$$

**Theorem 2.2.** *If $S$ is a $\sqrt{\epsilon/4}$ subspace embedding for $\mathtt{colspan}([A, b])$, then $\tilde{x}_{opt} = \text{argmin}_x \|SAx - Sb\|_2^2 = (SA)^+(Sb)$ is a $(1+\epsilon)$-approximate solution for the regression problem $\min_x \|Ax - b\|_2^2$.*

*Proof.* The proof goes along the line of Sarlos (2006). Let $A = U\Sigma V^T$ be the singular value decomposition of $A$. Define $x_{opt} = \text{argmin}_x \|Ax-b\|_2^2$. Define $\alpha$ such that $Ax_{opt} = U\alpha$ and $\beta$ such that $A\tilde{x}_{opt} - Ax_{opt} = U\beta$. Let $w = b - Ax_{opt}$. Let $\text{OPT} = \min_x \|Ax-b\|_2^2 = \|w\|_2^2$. We have $\text{colspan}(A, b) = \text{colspan}(U, w)$ and $U^T w = 0$. Let $S$ be a $\sqrt{\epsilon/4}$ subspace embedding of $\text{colspan}(A, b)$. We bound the cost of $\tilde{x}_{opt}$ as follows:

$$\begin{aligned} \|A\tilde{x}_{opt} - b\|_2^2 &= \|Ax_{opt} - b\|_2^2 + \|A\tilde{x}_{opt} - Ax_{opt}\|_2^2 \\ &\qquad \text{(Pythagorean Theorem)} \\ &= \text{OPT} + \|U\beta\|_2^2 \\ &= \text{OPT} + \|\beta\|_2^2 \end{aligned}$$

We get an upper-bound on $\|\beta\|_2$ in terms of OPT below.

$$\begin{aligned} \|\beta\|_2 - \|U^T S^T S U \beta\|_2 &\leq \|(I - U^T S^T S U)\beta\|_2 \\ &\leq \sqrt{\epsilon/4}\|\beta\|_2 \quad \text{(By Lemma 2.1)} \\ \implies \|\beta\|_2 &\leq \frac{\|U^T S^T S U \beta\|_2}{1 - \sqrt{\epsilon/4}} \end{aligned}$$

We show that $U^T S^T S U \beta = U^T S^T S w$ and use the fact that $S$ satisfies an approximate matrix multiplication property to bound $\|U^T S^T S w\|_2 = \|U^T S^T S U \beta\|_2$.

$$\begin{aligned} SU(\alpha+\beta) &= SA\tilde{x}_{opt} \\ &= SA((SA)^+ Sb)) \\ &= \Pi_{SA}(Sb) \\ &= \Pi_{SU}(Sb) \\ &\qquad \text{(Since colspan}(SA) = \text{colspan}(SU)) \\ &= \Pi_{SU}(S(w + U\alpha)) \\ &= SU\alpha + \Pi_{SU}(Sw) \end{aligned}$$

Hence,

$$\begin{aligned} SU\beta &= \Pi_{SU}(Sw) \\ \implies U^T S^T S U \beta &= U^T S^T S w. \end{aligned}$$

The last implication follows from the fact that for all $A, x : A^T \Pi_A(x) = A^T x$. Now,

$$\begin{aligned} \|U^T S^T S w\|_2 &= \|U^T S^T S w - U^T w\|_2 \quad \text{(Since } U^T w = 0) \\ &\leq \sqrt{\epsilon/4}\|U\|_2 \|w\|_2 \quad \text{(By Lemma 2.1)} \\ &= \sqrt{\epsilon/4 \cdot \text{OPT}}. \end{aligned}$$

Finally,

$$\begin{aligned}\|\beta\|^2 &\leq \frac{\|U^T S^T S U \beta\|^2}{(1-\sqrt{\epsilon/4})^2}\\ &\leq \frac{\epsilon/4}{(1-\sqrt{\epsilon/4})^2}\mathrm{OPT}\\ &\leq \epsilon\mathrm{OPT}.\end{aligned}$$

Therefore, $\|A\tilde{x}_{opt} - b\|^2 = \mathrm{OPT} + \|\beta\|^2 \leq (1+\epsilon)\mathrm{OPT}$. □

**Theorem 2.3.** *Given matrix $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, there exists a matrix $S$ which selects $O(d/\epsilon)$ rows of $A$ and scales them such that solution to the regression problem $\mathrm{argmin}_x \|SAx - Sb\|_2^2$ is a $(1+\epsilon)$-approximation to the regression problem $\|Ax - b\|_2^2$. This implies the existence of an $O(d/\epsilon)$-sized coreset.*

*Proof.* Applying BSS to the matrix $[A, b]$ with parameter $O(\sqrt{\epsilon})$ gives a selecting and rescaling matrix $S$ with $O(d/\epsilon)$ rows such that $S$ is a $\sqrt{\epsilon/4}$ subspace embedding for colspan$(A, b)$. By Theorem 2.2, we get that the solution to the regression problem $\min_x \|SAx - Sb\|_2^2$ is a $(1+\epsilon)$ approximate solution to the problem $\min_x \|Ax - b\|_2^2$. □

**Theorem 2.4.** *Given matrix $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, there exists a matrix $S$ which selects $O(d^2)$ rows of $A$ and scales them such that the solution to the regression problem $\mathrm{argmin}_x \|SAx - Sb\|_2^2$ is an optimal solution to $\mathrm{argmin}_x \|Ax - b\|_2^2$.*

*Proof.* The proof of this theorem is similar to that of (Maalouf et al., 2019), and is included here for completeness. Assume that the matrix $A$ is full rank. Let $a_i$ be the $i^{th}$ row of $A$ written as a column. Let $\tilde{a}_i \in \mathbb{R}^{d+1}$ be the vector $a_i$ appended with $b_i$. Consider the matrices $\tilde{a}_i\tilde{a}_i^T$ for $i = 1 \ldots n$. The matrix $(1/n)\sum_{i=1}^n \tilde{a}_i\tilde{a}_i^T$ lies in the convex hull of the matrices $\tilde{a}_i\tilde{a}_i^T$ for $i = 1 \ldots n$. By Caratheodory's theorem, there exists a set $\mathcal{S} \subseteq [n]$, $|\mathcal{S}| = O(d^2)$ and corresponding weights $w_i \geq 0$ for $i \in \mathcal{S}$, such that

$$\frac{1}{n}\sum_{i=1}^n \tilde{a}_i\tilde{a}_i^T = \sum_{j\in\mathcal{S}} w_j \tilde{a}_j\tilde{a}_j^T$$

We obtain the following relations from the above:

$$\sum_{i=1}^n a_i a_i^T = \sum_{j\in\mathcal{S}} (nw_j)\ a_j a_j^T$$

$$\sum_{i=1}^n b_i a_i = \sum_{j\in\mathcal{S}} (nw_j)\ b_j a_j$$

Let $s_1, s_2, \ldots, s_{|\mathcal{S}|} \in [n]$ be the elements of $\mathcal{S}$. Define a sampling and rescaling matrix $S$ as follows

$$S_{i,s_i} = \sqrt{nw_{s_i}}\ \ i = 1 \ldots |\mathcal{S}|$$

and the rest of the entries of $S$ are $0s$. Then,

$$\begin{aligned}&\operatorname*{argmin}_x \|SAx - Sb\|_2^2 = (A^TS^TSA)^{-1}(A^TS^TSb)\\ &= \left(\sum_{i=1}^{|\mathcal{S}|} S_{i,s_i}^2 a_{s_i} a_{s_i}^T\right)^{-1}\left(\sum_{i=1}^{|\mathcal{S}|} S_{i,s_i}^2 b_{s_i} a_{s_i}\right)\\ &= \left(\sum_{j\in\mathcal{S}} nw_j\ a_j a_j^T\right)^{-1}\left(\sum_{j\in\mathcal{S}} nw_j\ b_j a_j\right)\\ &= \left(\sum_{i=1}^n a_i a_i^T\right)^{-1}\left(\sum_{i=1}^n b_i a_i\right)\\ &= (A^TA)^{-1}(A^Tb)\\ &= \operatorname*{argmin}_x \|Ax - b\|_2^2\end{aligned}$$

□

## 2.2 Multiple Response Linear Regression

We now consider the problem of multiple response linear regression, where given matrices $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times d'}$, we find the solution of the following optimization problem

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

**Theorem 2.5.** *Given matrices $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times d'}$, if $S$ is a $\sqrt{\epsilon/4}$ subspace embedding for $\texttt{colspan}(A, B)$, then the solution to the optimization problem*

$$\tilde{X} = \operatorname*{argmin}_{X\in\mathbb{R}^{d\times d'}} \|SAX - SB\|_F^2$$

*is a $1+\epsilon$ approximate solution to the multiple response regression problem on matrices $A, B$, i.e.,*

$$\|A\tilde{X} - B\|_F^2 \leq (1+\epsilon)\min_{X\in\mathbb{R}^{d\times d'}} \|AX - B\|_F^2.$$

*Such a matrix $S$ with $O((d+d')/\epsilon)$ rows can be obtained using BSS.*

*Proof.* Let $x_i$ denote the $i$-th column of $X$ and $b_i$ be the $i$-th column of matrix $B$. Then the multiple response linear regression can be written as

$$\min_{x_1, x_2, \ldots, x_{d'}} \sum_i \|Ax_i - b_i\|_2^2.$$

These are $d'$ independent single response linear regression problems and given that $S$ is a $\sqrt{\epsilon/4}$ subspace embedding for $\texttt{colspan}(A, B)$, we get that for all $i = 1 \ldots d'$, $S$ is a $\sqrt{\epsilon/4}$ subspace embedding for $\texttt{colspan}(A, b_i)$. From Theorem 2.2, $\tilde{x}_i = \min_x \|SAx - Sb_i\|_2^2$ is a $1+\epsilon$ approximate solution to the regression problem on $A$ and $b_i$ and hence,

$$\sum_{i=1}^{d'} \|SA\tilde{x}_i - b_i\|_2^2 \leq \sum_{i=1}^{d'} (1+\epsilon)\min_{x_i} \|Ax_i - b_i\|_2^2.$$

So, the matrix $\tilde{X}$ having $i$th column equal to $\tilde{x}_i$ is a $(1+\epsilon)$ approximate solution for the regression problem on $(A, B)$. Thus,

$$\|A\tilde{X} - B\|_F^2 \leq (1+\epsilon) \min_X \|AX - B\|_F^2. \qquad \square$$

# 3 Upper Bounds for Ridge Regression - Statistical Dimension

In this section, we extend our results to the case of ridge regression and present coresets of size $O((\mathtt{sd}_\lambda + d')/\epsilon)$. We use approximate matrix product techniques of Cohen et al. (2016) to obtain the bounds in terms of statistical dimension.

**Theorem 3.1.** *Given matrices $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{n\times d'}$ and $\lambda \geq 0$, there exists a matrix $S$ which selects and scales $O((\mathtt{sd}_\lambda + d')/\epsilon)$ rows of $A$ such that solution to the ridge regression problem*

$$\min_X \|SAX - SB\|_2^2 + \lambda\|X\|_F^2$$

*is a $(1+\epsilon)$ approximate solution to the ridge regression problem*

$$\min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2.$$

*Proof.* Consider the matrix $\hat{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I_d \end{bmatrix}$ and $\hat{B} = \begin{bmatrix} B \\ 0_{d\times d'} \end{bmatrix}$. Let $\mathcal{A}$ be a matrix with orthonormal columns such that $\mathtt{range}(\mathcal{A}) = \mathtt{range}([\hat{A}\ \hat{B}])$ and the first $d$ columns of $\mathcal{A}$ are a basis for $\mathtt{colspan}(\hat{A})$. Let

$$\mathcal{A} = \begin{bmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{bmatrix} = \begin{bmatrix} U_1 & U_1' \\ U_2 & U_2' \end{bmatrix}$$

where $\mathcal{U}_1 \in \mathbb{R}^{n\times(d+d')}$, $\mathcal{U}_2 \in \mathbb{R}^{d\times(d+d')}$, $U_1 \in \mathbb{R}^{n\times d}$, $U_1' \in \mathbb{R}^{n\times d'}, U_2 \in \mathbb{R}^{d\times d}$ and $U_2 \in \mathbb{R}^{d\times d'}$. We have $\|\mathcal{U}_1\|_2^2 = \|[U_1\ U_1']\|_2^2 \leq 1$ and using Lemma 1.6 we get

$$\|\mathcal{U}_1\|_F^2 = \|[U_1\ U_1']\|_F^2 = \|U_1\|_F^2 + \|U_1'\|_F^2 \leq \mathtt{sd}_\lambda(A) + d'.$$

By Corollary 1 of (Cohen et al., 2016), we can obtain a *selecting* and *scaling* matrix $S$ with $O((\mathtt{sd}_\lambda + d')/(\epsilon/16)) = O((\mathtt{sd}_\lambda + d')/\epsilon)$ rows such that

$$\begin{aligned} \|\mathcal{U}_1^T S^T S \mathcal{U}_1 - \mathcal{U}_1^T \mathcal{U}_1\|_2^2 &\leq \frac{\epsilon}{16}\left(\|\mathcal{U}_1\|_2^2 + \frac{\|\mathcal{U}_1\|_F^2}{\mathtt{sd}_\lambda(A) + d'}\right)^2 \\ &\leq \frac{\epsilon}{16}\left(1 + \frac{\mathtt{sd}_\lambda(A) + d'}{\mathtt{sd}_\lambda(A) + d'}\right)^2 \\ &= \epsilon/4 \end{aligned}$$

Consider the *selecting* and *scaling* matrix

$$\mathcal{S} = \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix}$$

We have $\|\mathcal{A}^T\mathcal{S}^T\mathcal{S}\mathcal{A} - \mathcal{A}^T\mathcal{A}\|_2^2 = \|\mathcal{U}_1^T S^T S\mathcal{U}_1 + \mathcal{U}_2^T\mathcal{U}_2 - \mathcal{U}_1^T\mathcal{U}_1 - \mathcal{U}_2^T\mathcal{U}_2\|_2^2 = \|\mathcal{U}_1^T S^T S\mathcal{U}_1 - \mathcal{U}_1^T\mathcal{U}_1\|_2^2 \leq \epsilon/4$. Hence, $\mathcal{S}$ is a $\sqrt{\epsilon/4}$ subspace embedding for $\mathtt{range}(\mathcal{A}) = \mathtt{range}[\hat{A}\ \hat{B}]$. By Theorem 2.5, we have that the solution to the regression problem

$$\begin{aligned} \min_X \|\mathcal{S}\hat{A}X - \mathcal{S}\hat{B}\|_F^2 &= \min_X \| \begin{bmatrix} SA \\ \sqrt{\lambda} I \end{bmatrix} X - \begin{bmatrix} SB \\ 0 \end{bmatrix} \|_F^2 \\ &= \min_X \|SAX - SB\|_F^2 + \lambda\|X\|_F^2 \end{aligned}$$

is a $(1+\epsilon)$ approximate solution to

$$\begin{aligned} &\min_X \|\hat{A}X - \hat{B}\|_F^2 = \min_X \| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} X - \begin{bmatrix} B \\ 0 \end{bmatrix} \|_F^2 \\ &= \min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2 \end{aligned} \qquad \square$$

# 4 Deterministic Communication Protocol for Ridge Regression

## 4.1 Communication Model

We consider the communication model in which there are $s$ servers and there is a central coordinator which can communicate with every server. All communication occurs through two-way communication channels between the servers and the coordinator. The coordinator initiates the communication protocol and always decides who speaks next. This model is simpler to analyze and can simulate the arbitrary peer-to-peer communication model with a communication complexity of at most twice that of the peer-to-all model, by instead of having server A talk directly to server B, having server A forward its message through the coordinator. We must also add $\log s$ bits per message to tell the coordinator who to forward the message to.

## 4.2 Ridge Regression in the Distributed Setting

Consider the setting of ridge regression in a row-partition distributed setting. Let there be $s$ servers with matrices $A_1, A_2, \ldots, A_s$ and corresponding label matrices $B_1, B_2, \ldots, B_s$, respectively. Let $A$ be the matrix obtained by stacking $A_1, A_2, \ldots, A_s$ and $B$ be the matrix obtained by stacking $B_1, B_2, \ldots, B_s$. Assume that $\epsilon$ and all the entries are multiples of $1/\text{poly}(nd)$ and are upper bounded by $\text{poly}(nd)$. Therefore by multiplying all the entries by $\text{poly}(nd)$, we can assume that all the entries are integers and are upper bounded by $\text{poly}(nd)$ and hence each entry takes $O(\log(nd))$ bits to encode. This assumption also ensures that all the weights evaluated can be rounded to be encoded using $O(\log(nd)) + O(\log(1/\epsilon))$ bits. We call $O(\log(nd))$ bits a *word*. Let there be a central coordinator each server can communicate with. We would like to compute a $(1+\epsilon)$ approximate solution to the following optimization problem while minimizing the communication required

$$\begin{aligned}&\min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2\\ &= \min_X \left(\sum_{i=1}^{s} \|A_iX - B_i\|_F^2\right) + \lambda\|X\|_F^2\end{aligned}$$

**Theorem 4.1.** *If rows of matrix $A \in \mathbb{R}^{n\times d}$ are partitioned among $s$ servers and corresponding rows of $B \in \mathbb{R}^{n\times d'}$ are partitioned too, then there is a deterministic communication protocol using*

$$O(\frac{s\cdot t\cdot(\min{(s\cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A))} + d')}{\epsilon}) \text{ words.}$$

*where $t$ is the maximum number of non-zero entries in a row of $[A, B]$.*

*Proof.* For each $i$, define the following matrices

$$\hat{A}_i = \begin{bmatrix} A_i \\ \sqrt{\lambda/s}I_d\end{bmatrix}$$
$$\hat{B}_i = \begin{bmatrix} B_i \\ 0\end{bmatrix}$$

Let $\mathcal{A}_i$ be a matrix with orthonormal columns such that $\mathtt{range}(\mathcal{A}_i) = \mathtt{range}([\hat{A}_i \;\; \hat{B}_i])$. From the proof of Theorem 3.1, we obtain a $\sqrt{\epsilon/4}$ subspace embedding $\mathcal{S}_i$ of the form $\begin{bmatrix} S_i & 0 \\ 0 & I\end{bmatrix}$ for $\mathtt{range}(\mathcal{A}_i) = \mathtt{range}([\hat{A}_i \;\; \hat{B}_i])$. The matrix $S_i$ selects and scales $O((\mathtt{sd}_{\lambda/s}(A_i) + d')/\epsilon)$ rows of $A_i$ and $B_i$. Now, we have that the matrix

$$\mathcal{S} = \mathtt{diag}(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_s)$$

is a $\sqrt{\epsilon/4}$ subspace-embedding for

$$\mathtt{range}(\begin{bmatrix}\mathcal{A}_1\\ \vdots \\ \mathcal{A}_s\end{bmatrix}) = \mathtt{range}(\begin{bmatrix}\hat{A}_1 & \hat{B}_1\\ \vdots & \vdots \\ \hat{A}_s & \hat{B}_s\end{bmatrix}).$$

From Theorem 2.5, we obtain that solution to the optimization problem

$$\begin{aligned}&\min_X \|\mathcal{S}\begin{bmatrix}\hat{A}_1\\ \vdots \\ \hat{A}_s\end{bmatrix} X - \mathcal{S}\begin{bmatrix}\hat{B}_1\\ \vdots \\ \hat{B}_s\end{bmatrix}\|_F^2\\ &= \min_X \sum_i \|\mathcal{S}_i\hat{A}_iX - \mathcal{S}_i\hat{B}_i\|_F^2\\ &= \min_X \sum_i (\|S_iA_ix - S_iB_i\|_F^2 + \frac{\lambda}{s}\|X\|_F^2)\\ &= \min_X (\sum_i \|S_iA_iX - S_iB_i\|_F^2) + \lambda\|X\|_F^2\\ &= \min_X \|SAX - b\|_2^2 + \lambda\|X\|_2^2\end{aligned}$$

where $S$ is defined as

$$S = \mathtt{diag}(S_1, S_2, \ldots, S_s)$$

is a $(1+\epsilon)$ approximate solution to the regression problem.

$$\begin{aligned}&\min_X \|\begin{bmatrix}\hat{A}_1\\ \vdots \\ \hat{A}_s\end{bmatrix} X - \begin{bmatrix}\hat{B}_1\\ \vdots \\ \hat{B}_s\end{bmatrix}\|_F^2\\ &= \min_X \sum_i \|\hat{A}_iX - \hat{B}_i\|_F^2\\ &= \min_X \sum_i (\|A_iX - B_i\|_F^2 + \frac{\lambda}{s}\|X\|_F^2)\\ &= \min_X (\sum_i \|A_iX - B_i\|_F^2) + \lambda\|X\|_F^2\\ &= \min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2.\end{aligned}$$

The communication protocol is as follows: the $i$-th server computes the *selecting* and *scaling* matrix $S_i$ as above and sends the the matrices $S_iA_i$ and $S_iB_i$ to the central server (also called the coordinator). The central server can now compute the solution to the problem $\min_X \sum_i \|S_iA_iX - S_iB_i\|_F^2 + \lambda\|X\|_F^2$ by standard techniques. The solution obtained is guaranteed to be a $(1+\epsilon)$ solution as shown above.

When each row of matrix $[A_i, B_i]$ has at most $t$ non-zero entries, the communication required is at most $O(s\cdot t\cdot \frac{\max_i \mathtt{sd}_{\lambda/s}(A_i) + d'}{\epsilon})$ words for the entries of the matrices and $O(s\cdot \frac{\max_i \mathtt{sd}_{\lambda/s}(A_i) + d'}{\epsilon})$ words for the weights. But, for all $i$, $\mathtt{sd}_{\lambda/s}(A_i) \le \mathtt{sd}_{\lambda/s}(A) \le \min(s\cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A))$. Hence, the communication complexity is $O(\frac{s\cdot t\cdot(\min{(s\cdot \mathtt{sd}_\lambda(A), \mathtt{rank}(A))} + d')}{\epsilon})$ words. □

## 5 Lower Bounds for Multi Response Ridge Regression

In this section, we give example matrices $A, B \in \mathbb{R}^{d/100\epsilon\times d}$, $\epsilon > 0$, $\lambda \ge 0$ such that $\mathtt{sd}_\lambda(A) = \Omega(d)$ and any *selecting* and *scaling* matrix $S$ needs at least $\Omega(d/\epsilon)$ rows for it to give a $1+\epsilon$ approximate solution.

**Theorem 5.1.** *For all $\epsilon$ such that $1 \le 1/100\epsilon \le d$ and $\lambda \le 1/4\epsilon$ there exist matrices $A, B \in \mathbb{R}^{d/100\epsilon\times d}$ for which any matrix $S$ that selects and rescales $k$ rows of $A$ and $B$ such that the solution to*

$$\min_X \|SAX - SB\|_F^2 + \lambda\|X\|_F^2$$

*is a $(1+\epsilon)$ approximation to*

$$\min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2$$

*has $k = \Omega(\mathtt{sd}_\lambda(A)/\epsilon)$ rows.*

*Proof.* Let the matrix $A$ be a block matrix where each block has dimensions $1/100\epsilon \times 1$. Define blocks on the diagonal of $A$ to be the vectors $1_{1/100\epsilon}$ and remaining entries of $A$ to be 0. The singular values of this matrix are all equal to $\sqrt{1/100\epsilon}$. For $\lambda \leq 1/4\epsilon$, $\mathtt{sd}_\lambda(A) \geq d/26$. Similarly, let $B$ be a block matrix, where each block has size $1/100\epsilon \times d$. So, the matrix $B$ is formed by stacking matrices $B_1, B_2, \dots, B_d$. Let each row of $B_i$ be a distinct unit vector in the standard basis for $\mathbb{R}^d$. We can choose $1/100\epsilon$ distinct standard basis vectors as $d \geq 1/100\epsilon$. For the block matrix $B_i$, define a set $H_i \subseteq [d]$ of integers $k$ such that $e_k$ is a row in $B_i$.

The problem $\min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2$ is equivalent to

$$\begin{aligned}&\min_{x_1, x_2, \dots, x_d} (\|1_{1/100\epsilon}x_1^T - B_1\|_F^2 + \lambda\|x_1\|_2^2) + \ldots + \\ &(\|1_{1/100\epsilon}x_d^T - B_d\|_F^2 + \lambda\|x_d\|_2^2) \qquad (5.1)\end{aligned}$$

and the above is equivalent to minimizing each of the problems independently.

We consider the problem

$$\min_{x_1} \|1_{1/100\epsilon}x_1^T - B_1\|_F^2 + \lambda\|x_1\|_2^2. \qquad (5.2)$$

Without loss of generality assume that $H_1 = [1/100\epsilon]$ i.e., rows of $B_1$ are the first $1/100\epsilon$ standard basis vectors of $\mathbb{R}^d$. Consider a matrix $S$ which selects $k$ rows of $[1_{1/100\epsilon}, B_1]$ and solves the following optimization problem

$$\min_{x_1} \|S1_{1/100\epsilon}x_1^T - SB_1\|_F^2 + \lambda\|x_1\|_2^2$$

We can assume that $S$ selects the top $k$ rows without loss of generality. Then the above problem is equivalent to

$$\begin{aligned}&\min_{y_1, y_2, \dots, y_d \in \mathbb{R}} \sum_{i=1}^{k} (w_i(1-y_i)^2 + (W - w_i)y_i^2 + \lambda y_i^2) \\ &+ \sum_{i=k+1}^{d} (\lambda + W)y_i^2\end{aligned}$$

where $w_i$ is the weight $S$ assigns to error in row $i$ and $W = \sum_{i=1}^k w_i$. By taking the partial derivative of the objective function with respect to $y_i$ and setting it to 0, we get that it is minimized when $y_i = w_i/(W+\lambda)$ for $i = 1 \dots k$ and $y_i = 0$ for $i = k+1 \dots d$. When this solution is used for the original ridge regression problem (5.2), the cost is

$$\begin{aligned}&\frac{1}{100\epsilon} - k \\ &+ \sum_{i=1}^{k} \left(1 - \frac{w_i}{W+\lambda}\right)^2 + \left(1 - \frac{1}{100\epsilon}\right)\left(\frac{w_i}{W+\lambda}\right)^2 \\ &+ \lambda\left(\frac{w_i}{W+\lambda}\right)^2\end{aligned}$$

For a fixed $W$, the cost is minimized when all $w_i$'s are equal and hence we set $w_i = W/k = b$ for some $b \geq 0$. The cost can now be written as

$$\begin{aligned}&\frac{1}{100\epsilon} - k \\ &+ k\left[\left(1 - \frac{b}{\lambda + kb}\right)^2 + \left(\frac{1}{100\epsilon} - 1\right)\left(\frac{b}{\lambda + bk}\right)^2\right. \\ &\left.+\lambda\left(\frac{b}{\lambda + bk}\right)^2\right] \\ &= \frac{1}{100\epsilon} - k + k\left[1 + c^2 - 2c + \left(\frac{1}{100\epsilon} - 1\right)c^2 + \lambda c^2\right] \\ &\quad \text{where } c = \frac{b}{\lambda + bk} \\ &= \frac{1}{100\epsilon} - k + k + k\left[\left(\frac{1}{100\epsilon} + \lambda\right)c^2 - 2c\right] \\ &= \frac{1}{100\epsilon} + k\left[\left(\frac{1}{100\epsilon} + \lambda\right)c^2 - 2c\right]. \qquad (5.3)\end{aligned}$$

This is minimized when $c = \frac{1}{\lambda + 1/100\epsilon}$ which is obtained when $b = \frac{\lambda}{\lambda + (1/100\epsilon - k)}$ (This is a valid setting of weights as $k \leq 1/100\epsilon$ and hence $b \geq 0$). The minimum value is hence equal to $1/100\epsilon - k/(\lambda + 1/100\epsilon)$. This is the least error we can get on (5.2) using any matrix $S$ which selects and re-scales $\leq k$ rows. Substituting $k = 1/100\epsilon$ we recover the OPT value for (5.2) which is equal to $1/100\epsilon - 1/(1 + 100\lambda\epsilon)$. Now, $1/100\epsilon - k/(\lambda + 1/100\epsilon)$ is $\leq (1+2\epsilon)\mathrm{OPT}$ iff

$$\begin{aligned}&\frac{1}{100\epsilon} - \frac{k}{\lambda + 1/100\epsilon} \leq (1+2\epsilon)\left(\frac{1}{100\epsilon} - \frac{1}{1+100\lambda\epsilon}\right) \\ \text{iff} \quad &\frac{1}{100\epsilon} - \frac{k}{\lambda + 1/100\epsilon} \leq \frac{1}{100\epsilon} + \frac{1}{50} - \frac{1+2\epsilon}{1+100\lambda\epsilon} \\ \text{iff} \quad &-k \leq \frac{\lambda + 1/100\epsilon}{50} - \frac{1+2\epsilon}{100\epsilon} \\ &= \frac{2\epsilon\lambda + 1/50 - 1 - 2\epsilon}{100\epsilon} \\ \text{iff} \quad &k \geq \frac{49/50 - 2\epsilon\lambda + 2\epsilon}{100\epsilon}.\end{aligned}$$

For any $\lambda \leq 1/4\epsilon$, this implies that $k \geq 1/400\epsilon$.

To get a $(1+\epsilon)$ approximate solution to (5.1), we need to solve at least $d/2$ sub-problems upto $(1+2\epsilon)$ approximation. Hence, the selecting matrix $S$ for the whole problem must select at least $d/2 \times 1/400\epsilon = \Omega(d/\epsilon) = \Omega(\mathtt{sd}_\lambda(A)/\epsilon)$ rows. □

This shows the $O(d^2)$ upper bound to construct covariance matrices using Caratheodory's theorem is tight.

**Acknowledgments:** The authors would like to thank support in part by the Office of Naval Research (ONR) grant N00014-18-1-2562.

## References

Zeyuan Allen Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and regret minimization beyond matrix multiplicative updates. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 237–245, 2015. doi: 10.1145/2746539.2746610. URL https://doi.org/10.1145/2746539.2746610.

Alexandr Andoni, Jiecao Chen, Robert Krauthgamer, Bo Qin, David P. Woodruff, and Qin Zhang. On sketching quadratic forms. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 311–319, 2016. doi: 10.1145/2840728.2840753. URL https://doi.org/10.1145/2840728.2840753.

Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regularized data fitting. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, pages 27:1–27:22, 2017. doi: 10.4230/LIPIcs.APPROX-RANDOM.2017.27. URL https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2017.27.

Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

Charles Carlson, Alexandra Kolla, Nikhil Srivastava, and Luca Trevisan. Optimal lower bounds for sketching graph cuts. *CoRR*, abs/1712.10261, 2017. URL http://arxiv.org/abs/1712.10261.

Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 663–695, 2019. URL http://proceedings.mlr.press/v99/chen19a.html.

Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 11:1–11:14, 2016. doi: 10.4230/LIPIcs.ICALP.2016.11. URL https://doi.org/10.4230/LIPIcs.ICALP.2016.11.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices I: approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006a. doi: 10.1137/S0097539704442684. URL https://doi.org/10.1137/S0097539704442684.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices II: computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006b. doi: 10.1137/S0097539704442696. URL https://doi.org/10.1137/S0097539704442696.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices III: computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1):184–206, 2006c.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011. doi: 10.1007/s00211-010-0331-6. URL https://doi.org/10.1007/s00211-010-0331-6.

Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. *SIAM J. Comput.*, 47(6):2315–2336, 2018. doi: 10.1137/16M1061850. URL https://doi.org/10.1137/16M1061850.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. *CoRR*, abs/1906.04705, 2019. URL http://arxiv.org/abs/1906.04705.

Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. doi: 10.1561/2200000035. URL https://doi.org/10.1561/2200000035.

T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152, Oct 2006. doi: 10.1109/FOCS.2006.37.

Santosh S. Vempala, Ruosong Wang, and David P. Woodruff. The communication complexity of optimization. *CoRR*, abs/1906.05832, 2019. URL http://arxiv.org/abs/1906.05832.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014. doi: 10.1561/0400000060. URL https://doi.org/10.1561/0400000060.