

Near-optimal Coresets For Least-Squares Regression

Christos Boutsidis*

Petros Drineas†

Malik Magdon-Ismail‡

June 25, 2013

Abstract

We study (constrained) least-squares regression as well as multiple response least-squares regression and ask the question of whether a subset of the data, a *coreset*, suffices to compute a good approximate solution to the regression. We give deterministic, low order polynomial-time algorithms to construct such coresets with approximation guarantees, together with lower bounds indicating that there is not much room for improvement upon our results.

1 Introduction

Linear regression is an important technique in data analysis [18]. Research in the area ranges from numerical techniques [1] to robustness of the prediction error to noise (e.g., using feature selection [13]). We ask whether it is possible to efficiently identify a small subset of the data that contains all the essential information of a learning problem. Such a subset is called a “coreset”. We show that the answer is yes, for linear regression. Such a coreset is analogous to the support vectors in support vector machines [9]. Such coresets contain the meaningful or important points in the data and can be used to find good approximate solutions to the full problem by solving a (much) smaller problem. When the constraints are complex (e.g., non-convex constraints), solving a much smaller regression problem could be a significant saving [12].

We present coreset constructions for constrained regression (both simple and multiple response), as well as lower bounds for the size of coresets that achieve certain accuracy. In addition to potential computational savings, a coreset identifies the important core of a machine learning problem and is of considerable interest in applications with huge data where incremental approaches are necessary (for example chunking) and applications where the data is distributed and bandwidth is costly (hence communicating only the essential data is imperative [15]).

Our first contribution is a deterministic, polynomial-time algorithm for constructing a coreset for arbitrarily constrained linear regression. Let k be the “effective dimension” of the data (the rank of the data matrix) and let $\epsilon > 0$ be the desired accuracy parameter. Our algorithm constructs a coreset of size $O(k/\epsilon^2)$, which achieves a $(1 + \epsilon)$ -relative error performance guarantee. In other words, solving the regression problem on the coreset results in a solution which fits all the data with an error which is at most $(1 + \epsilon)$ worse than the best possible fit to all the data. We extend our results to the setting of multiple response regression using more sophisticated techniques. Our proofs are based on two sparsification tools from linear algebra [2, 7], which may be of general interest to the machine learning community, and we discuss these in some detail.

*Mathematical Sciences Department, IBM T.J. Watson Research Center. Email: cbouts@us.ibm.com.

†Computer Science Department, Rensselaer Polytechnic Institute. Email: drinep@cs.rpi.edu.

‡Computer Science Department, Rensselaer Polytechnic Institute. Email: magdon@cs.rpi.edu.

1.1 Problem Setup

Assume the usual setting with n data points $(\mathbf{z}_1, y_1), \dots, (\mathbf{z}_n, y_n)$; $\mathbf{z}_i \in \mathbb{R}^d$ are feature vectors (which could have been obtained by applying a non-linear feature transform to raw data) and $y_i \in \mathbb{R}$ are targets (responses). The linear regression problem asks to determine a vector $\mathbf{x}_{opt} \in \mathcal{D} \subseteq \mathbb{R}^d$ that minimizes

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n w_i (\mathbf{z}_i^T \mathbf{x} - y_i)^2,$$

over $\mathbf{x} \in \mathcal{D}$, where $w_i \in \mathbb{R}$ are positive weights. So, $\mathcal{E}(\mathbf{x}_{opt}) \leq \mathcal{E}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{D}$. The domain \mathcal{D} represents the constraints on the solution, e.g., in non-negative least squares (NNLS) [16, 3], $\mathcal{D} = \mathbb{R}_+^d$, the nonnegative orthant. Our results hold for arbitrary \mathcal{D} .

A coreset of size $r < n$ is a subset of the data points, $(\mathbf{z}_{i_1}, y_{i_1}), \dots, (\mathbf{z}_{i_r}, y_{i_r})$. The coreset regression problem considers the squared error on the coreset with a (possibly) different set of weights $s_j > 0$,

$$\tilde{\mathcal{E}}(\mathbf{x}) = \sum_{j=1}^r s_j (\mathbf{z}_{i_j}^T \mathbf{x} - y_{i_j})^2.$$

Suppose that $\tilde{\mathcal{E}}$ is minimized at $\tilde{\mathbf{x}}_{opt} \in \mathcal{D} \subseteq \mathbb{R}^d$, so $\tilde{\mathcal{E}}(\tilde{\mathbf{x}}_{opt}) \leq \tilde{\mathcal{E}}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{D}$. Such a coreset is of interest if, for some set of weights s_j , $\tilde{\mathbf{x}}_{opt}$ is nearly as good as \mathbf{x}_{opt} for the original regression problem on all the data. That is, for some small $\epsilon > 0$,

$$\mathcal{E}(\mathbf{x}_{opt}) \leq \mathcal{E}(\tilde{\mathbf{x}}_{opt}) \leq (1 + \epsilon)\mathcal{E}(\mathbf{x}_{opt}).$$

The algorithm which constructs the coreset should also provide the weights \mathbf{s}_j . For the remainder of the paper, we switch to an equivalent matrix formulation of the problem. (See Appendix for linear algebra background.)

1.1.1 Matrix Formulation

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be the data matrix whose rows are the weighted data points $\sqrt{w_i} \mathbf{z}_i^T$; and let $\mathbf{b} \in \mathbb{R}^n$ be the similarly weighted target vector, $b_i = \sqrt{w_i} y_i$, where for $i = 1, \dots, n$, b_i denotes the i th element of $\mathbf{b} \in \mathbb{R}^n$. The effective dimension of the data can be measured by the rank of \mathbf{A} ; let $k = \text{rank}(\mathbf{A})$. Our results hold for arbitrary $n > d$, however, in most applications, $n \gg d$ and $\text{rank}(\mathbf{A}) \approx d$. We can rewrite the squared error as $\mathcal{E}(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, so,

$$\mathbf{x}_{opt} \in \underset{\mathbf{x} \in \mathcal{D}}{\text{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \quad (1)$$

A coreset of size $r < n$ is a subset $\mathbf{C} \in \mathbb{R}^{r \times d}$ of the rows of \mathbf{A} and the corresponding elements $\mathbf{b}_c \in \mathbb{R}^r$ of \mathbf{b} . Let $\mathbf{D} \in \mathbb{R}^{r \times r}$ be a positive diagonal matrix for the coreset regression (the weights s_j of the coreset regression will depend on \mathbf{D}). The weighted squared error on the coreset is given by

$$\tilde{\mathcal{E}}(\mathbf{x}) = \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2,$$

so the coreset regression seeks $\tilde{\mathbf{x}}_{opt}$ defined by

$$\tilde{\mathbf{x}}_{opt} \in \underset{\mathbf{x} \in \mathcal{D}}{\text{argmin}} \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2.$$

We say that such a coreset is an $(1 + \epsilon)$ -coreset if the solution obtained by fitting the coreset data is almost optimal for all the data. Formally,

$$\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq (1 + \epsilon)\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

Type of Regression	Approximation Ratio	
Constrained single-response	$1 + O(\sqrt{k/r})$	[Eqn. (1), Thm. 1]
Multi-objective	$1 + O(\sqrt{k/r})$	[Eqn. (3), Thm. 4]
Constrained multiple-response (Frobenius)	$1 + O(\sqrt{k\omega/r})$	[Eqns. (4), (5)]
Unconstrained multiple-response (Spectral)	$2 + O(\sqrt{\omega/r} + \omega/r + \sqrt{k/r})$	[Eqn. (6), Thm. 6]
Unconstrained multiple-response (Frobenius)	$2 + O(\sqrt{k/r})$	[Eqn. (6), Thm. 7]
Unconstrained multiple-response (b -agnostic)	$O(n/r)$	[Eqn. (6), Thm. 12]

Table 1: Summary of our results for coreset construction in linear regression. In all cases, our algorithms are *deterministic* and construct a coreset of size r . The approximation ratios are values β such that $\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|/\|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\| \leq \beta$. In the first row in the table, \mathbf{X}_{opt} , $\tilde{\mathbf{X}}_{opt}$, and \mathbf{B} are vectors. NOTATION: n is the number of data points of dimension $d < n$; k is the rank of the matrix whose rows correspond to the n data points; r is the size of the coreset, $k < r < n$; $\omega \geq 1$ is the number of “response” vectors in multiple-response regression (in the last four rows in the table \mathbf{X}_{opt} , $\tilde{\mathbf{X}}_{opt}$, and \mathbf{B} have ω columns).

1.2 Our contributions

In this section, we discuss our main results for various formulations of linear regression (also summarized in Table 1). In the next section we present the relevant algorithms and proofs.

1.2.1 Constrained Linear Regression (Section 2)

Our main result for constrained simple regression is Theorem 1, which describes a *deterministic* polynomial time algorithm that constructs a $(1+\epsilon)$ -coreset of size $O(k/\epsilon^2)$. Prior to our work, the best result achieving comparable relative error performance guarantees is Theorem 1 of [6] for constrained regression, and the work of [11] for unconstrained regression. Both of these prior results construct coresets of size $O(k \log k/\epsilon^2)$ and they are randomized, so, with some probability, the fit on all the data can be arbitrarily bad (despite the coreset being a logarithmic factor larger). Our methods have comparable, low order polynomial running times and provide deterministic guarantees. The results in [11] and [6] were achieved using the matrix concentration results in [17]. However, these concentration bounds break unless the coreset size is $\Omega(k \log k/\epsilon^2)$.

We extend our results to multiple response regression, where the target is a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ with $\omega \geq 1$. Each column of \mathbf{B} is a separate target (or response) that we wish to predict. We seek to minimize $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|$ over all $\mathbf{X} \in \mathcal{D} \subseteq \mathbb{R}^{d \times \omega}$. Multiple response regression has numerous applications, but is perhaps most common in multivariate time series analysis; see for example [14, 8]. To illustrate, consider prediction of time series data: let $\mathbf{Z} \in \mathbb{R}^{(n+1) \times d}$ be a set of d time series, where each column is a time series with $n+1$ time steps; we wish to predict time step $t+1$ from time step t . Let \mathbf{A} contain the first n rows of \mathbf{Z} and let \mathbf{B} contain the last n rows. Then, we seek \mathbf{X} that minimizes $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{\xi}$ under some norm ξ , which is exactly the multiple response regression problem. In our work, we consider the spectral ($\xi = 2$) and Frobenius ($\xi = F$) norms.

1.2.2 Multi-Objective Regression (Section 3.1)

An important variant of multiple response regression is the so-called multi-objective regression. Let

$$\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_\omega] \in \mathbb{R}^{n \times \omega},$$

where we explicitly identify each column in \mathbf{B} as a target response $\mathbf{b}_j \in \mathbb{R}^n$ where $j \in \{1, 2, \dots, \omega\}$. We seek to simultaneously fit multiple target vectors with the same \mathbf{x} , i.e., to simultaneously minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}_j\|_2^2$. This is common when the goal is to trade off different quality criteria simultaneously. Writing $\mathbf{X} = [\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$ (ω copies of $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$), we consider minimizing $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F$, which is equivalent to multiple regression with a strong constraint on \mathbf{X} . We present results for coreset constructions for the Frobenius-norm multi-objective regression problem in Theorem 4, which describes a deterministic algorithm to construct $(1 + \epsilon)$ -coresets of size $O(k/\epsilon^2)$, where $k = \text{rank}(\mathbf{A})$. Theorem 4 emerges by applying Theorem 1 after converting the Frobenius-norm multi-objective regression problem to a simple response regression problem.

1.2.3 Arbitrarily-Constrained Multiple-Response Regression (Section 3.2)

Using the same approach, converting the problem to a single response regression, we construct a $(1 + \epsilon)$ -coreset for Frobenius-norm arbitrarily-constrained regression in Section 3.2. The coreset size in this case is $O(k\omega/\epsilon^2)$.

1.2.4 Unconstrained Multiple-Response Regression (Section 4)

In Section 4, we consider coresets for unconstrained multiple-response regression for both the spectral and Frobenius norms. The sizes of the coresets are smaller than the constrained case, and our main results are presented in Theorems 6 and 7. Theorem 6 presents a $(2 + \epsilon)$ -coreset of size $O((k + \omega)/\epsilon^2)$ for spectral norm regression, while Theorem 7 presents a $(2 + \epsilon)$ -coreset of size $O(k/\epsilon^2)$ for Frobenius norm regression.

1.2.5 Lower Bounds (Section 5)

Finally, in Section 5, we present lower bounds on coreset sizes. In the single response regression setting, we note that our algorithms need to look at the target vector \mathbf{b} . We show that this is unavoidable, by arguing that no \mathbf{b} -agnostic deterministic coreset construction algorithm can construct coresets which are small (Theorem 13). We also present similar results for \mathbf{b} -agnostic randomized coreset constructions (Theorem 14).

Then, we present lower bounds on the size of coresets for spectral and Frobenius norm multiple response regression that apply in the general, non \mathbf{b} -agnostic, setting (Theorems 15 and 16).

2 Constrained Linear Regression

We define constrained linear regression as follows: given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{D} \subseteq \mathbb{R}^d$, we seek $\mathbf{x}_{opt} \in \mathcal{D}$ for which $\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, for all $\mathbf{x} \in \mathcal{D}$ (the domain \mathcal{D} represents the constraints on \mathbf{x} and can be arbitrary). To construct a coreset $\mathbf{C} \in \mathbb{R}^{r \times d}$ (i.e., \mathbf{C} consists of r rows of \mathbf{A}) and $\mathbf{b}_c \in \mathbb{R}^r$ (i.e., \mathbf{b}_c consists of r elements of \mathbf{b}), we introduce *sampling* and *rescaling* matrices \mathbf{S} and \mathbf{D} respectively. More specifically, we define the *row-sampling matrix* $\mathbf{S} \in \mathbb{R}^{r \times n}$ whose rows are basis vectors $\mathbf{e}_{i_1}^T, \dots, \mathbf{e}_{i_r}^T$. Our coreset \mathbf{C} is now equal to $\mathbf{C} = \mathbf{S}\mathbf{A}$; clearly, \mathbf{C} is a matrix whose rows are the rows of \mathbf{A} corresponding to indices i_1, \dots, i_r . Similarly, $\mathbf{b}_c = \mathbf{S}\mathbf{b}$ contains the corresponding elements of the target vector. Next,

let $\mathbf{D} \in \mathbb{R}^{r \times r}$ be a positive diagonal rescaling matrix and define the \mathbf{D} -weighted regression problem on the coreset as follows:

$$\tilde{\mathbf{x}}_{opt} \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{D}} \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2 = \operatorname{argmin}_{\mathbf{x} \in \mathcal{D}} \|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2. \quad (2)$$

In the above, the operator $\mathbf{D}\mathbf{S}$ first samples and then rescales rows of \mathbf{A} and \mathbf{b} . Theorem 1 is the main result in this section and presents a deterministic algorithm to select a coreset by constructing \mathbf{D} and \mathbf{S} .

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $r > k + 1$.

Output: sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.

- 1: Compute the SVD of $\mathbf{Y} = [\mathbf{A}, \mathbf{b}]$. Let $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times \ell}$, $\mathbf{\Sigma} \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{V} \in \mathbb{R}^{(d+1) \times \ell}$, with $\ell \leq k + 1$ (the rank of \mathbf{Y}).
- 2: **Return** $[\mathbf{D}, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$ (see Lemma 2)

Algorithm 1: Deterministic coreset construction for constrained linear regression.

Theorem 1. *Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{D} \subseteq \mathbb{R}^d$, Algorithm 1 constructs matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ (for any $r > k + 1$) such that $\tilde{\mathbf{x}}_{opt}$ of Eqn. (2) satisfies*

$$\frac{\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2}{\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2} \leq \frac{r + k + 1 + 2\sqrt{r(k+1)}}{r + k + 1 - 2\sqrt{r(k+1)}} = 1 + 4\sqrt{\frac{k}{r}} + o\left(\sqrt{k/r}\right).$$

The running time of the proposed algorithm is $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}]}) + O(rnk^2)$, where $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}]})$ is the time needed to compute the left singular vectors of the matrix $[\mathbf{A}, \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$.

For any $0 < \epsilon < 1$, we can set $r = k/\epsilon^2$ to get an approximation ratio roughly equal to $1 + 4\epsilon$. This result considerably improves the result in [6], which needs $r = O(k \log k/\epsilon^2)$ to achieve the same approximation ratio. Additionally, our bound is deterministic, whereas the bound in [6] fails with constant probability. [6] also requires an SVD computation in the first step, so its running time is comparable to ours.

In order to prove the above theorem, we need a linear algebraic sparsification result from [2], specifically Theorem 3.1 in [2], which we restate using our notation (we present the corresponding algorithm below).

Lemma 2 (Single-set Spectral Sparsification [2]). *Given $\mathbf{U} \in \mathbb{R}^{n \times \ell}$ satisfying $\mathbf{U}^T \mathbf{U} = \mathbf{I}_\ell$ and $r > \ell$, we can deterministically construct sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ such that, for all $\mathbf{y} \in \mathbb{R}^\ell$:*

$$\left(1 - \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}\|_2^2 \leq \|\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}\|_2^2 \leq \left(1 + \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}\|_2^2.$$

The algorithm runs in $O(rn\ell^2)$ time and we denote it as $[\mathbf{D}, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$.

Proof. (of Theorem 1) Let $\mathbf{Y} = [\mathbf{A}, \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$ and compute its SVD: $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Let ℓ be the rank of \mathbf{Y} ($\ell \leq k + 1$, since $\operatorname{rank}(\mathbf{A}) = k$) and note that $\mathbf{U} \in \mathbb{R}^{n \times \ell}$, $\mathbf{\Sigma} \in \mathbb{R}^{\ell \times \ell}$, and $\mathbf{V} \in \mathbb{R}^{(d+1) \times \ell}$. Let $[\mathbf{D}, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$ and define $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^\ell$ as follows:

$$\mathbf{y}_1 = \mathbf{\Sigma}\mathbf{V}^T \begin{bmatrix} \mathbf{x}_{opt} \\ -1 \end{bmatrix}, \quad \text{and} \quad \mathbf{y}_2 = \mathbf{\Sigma}\mathbf{V}^T \begin{bmatrix} \tilde{\mathbf{x}}_{opt} \\ -1 \end{bmatrix}.$$

Note that $\mathbf{U}\mathbf{y}_1 = \mathbf{A}\mathbf{x}_{opt} - \mathbf{b}$, $\mathbf{U}\mathbf{y}_2 = \mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}$, $\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}_1 = \mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b})$, and $\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}_2 = \mathbf{D}\mathbf{S}(\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b})$. We will bound $\|\mathbf{U}\mathbf{y}_2\|_2$ in terms of $\|\mathbf{U}\mathbf{y}_1\|_2$:

$$\left(1 - \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}_2\|_2^2 \stackrel{(a)}{\leq} \|\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}_2\|_2^2 \stackrel{(b)}{\leq} \|\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}_1\|_2^2 \stackrel{(c)}{\leq} \left(1 + \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}_1\|_2^2.$$

Input: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]^\top \in \mathbb{R}^{n \times \ell}$ with $\mathbf{u}_i \in \mathbb{R}^\ell$ and $r > \ell$.

Output: Sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.

1: Initialize $\mathbf{A}_0 = \mathbf{0}_{\ell \times \ell}$, $\mathbf{S} = \mathbf{0}_{r \times n}$, and $\mathbf{D} = \mathbf{0}_{r \times r}$.

2: Set constants $\delta_L = 1$ and $\delta_U = (1 + \ell/r) \left(1 - \sqrt{\ell/r}\right)^{-1}$.

3: **for** $\tau = 0$ **to** $r - 1$ **do**

4: Let $L_\tau = \tau - \sqrt{r\ell}$; $U_\tau = \delta_U \left(\tau + \sqrt{\ell r}\right)$.

5: Pick index $i_\tau \in \{1, 2, \dots, n\}$ and number $t_\tau > 0$ (see Section 2.1 for the definition of U, L):

$$U(\mathbf{u}_{i_\tau}, \delta_U, \mathbf{A}_\tau, U_\tau) \leq \frac{1}{t_\tau} \leq L(\mathbf{u}_{i_\tau}, \delta_L, \mathbf{A}_\tau, L_\tau).$$

6: Update $\mathbf{A}_{\tau+1} = \mathbf{A}_\tau + t_\tau \mathbf{u}_{i_\tau} \mathbf{u}_{i_\tau}^\top$; and set $\mathbf{S}_{\tau+1, i_\tau} = 1$, $\mathbf{D}_{\tau+1, \tau+1} = 1/\sqrt{t_\tau}$.

7: **end for**

8: Multiply all the weights in \mathbf{D} by $\sqrt{r^{-1} \left(1 - \sqrt{\ell/r}\right)}$.

9: **Return:** \mathbf{S} and \mathbf{D} .

Algorithm 2: SimpleSampling (Lemma 2)

(a) and (c) follow from Lemma 2; (b) follows from the optimality of $\tilde{\mathbf{x}}_{opt}$ for the coresets regression in Eqn. (2). Using $\ell \leq k + 1$ and manipulating the above expression concludes the proof of the theorem. The running time of the algorithm is equal to the time needed to compute \mathbf{U} and the time needed to run the algorithm of Lemma 2 with $\ell \leq k + 1$. \blacksquare

2.1 Single-set Spectral Sparsification Algorithm (Lemma 2)

We now discuss in more detail the sparsification algorithm of Lemma 2. We present the corresponding algorithm as Algorithm 6. Our notation deviates from the original in [2]; we employ our own presentation of the corresponding algorithm in [7]. Algorithm 6 is a greedy technique that selects columns one at a time. To describe the algorithm in more detail, it is convenient to view the input matrix as a set of n column vectors,

$$\mathbf{U}^\top = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n],$$

with $\mathbf{u}_i \in \mathbb{R}^\ell$ ($i = 1, \dots, n$). Given ℓ and $r > \ell$, introduce the iterator $\tau = 0, 1, 2, \dots, r - 1$, and define the parameter $L_\tau = \tau - \sqrt{r\ell}$. For a square symmetric matrix $\mathbf{A} \in \mathbb{R}^{\ell \times \ell}$ with eigenvalues $\lambda_1, \dots, \lambda_\ell$, vector $\mathbf{u} \in \mathbb{R}^\ell$ and scalar $L \in \mathbb{R}$, define

$$\phi(L, \mathbf{A}) = \sum_{i=1}^{\ell} \frac{1}{\lambda_i - L},$$

and let $L(\mathbf{u}, \delta_L, \mathbf{A}, L)$ be defined as

$$L(\mathbf{u}, \delta_L, \mathbf{A}, L) = \frac{\mathbf{u}^\top (\mathbf{A} - L' \mathbf{I}_\ell)^{-2} \mathbf{u}}{\phi(L', \mathbf{A}) - \phi(L, \mathbf{A})} - \mathbf{u}^\top (\mathbf{A} - L' \mathbf{I}_\ell)^{-1} \mathbf{u},$$

where

$$L' = L + \delta_L = L + 1.$$

Similarly, for a square symmetric matrix $\mathbf{A} \in \mathbb{R}^{\ell \times \ell}$ with eigenvalues $\lambda_1, \dots, \lambda_\ell$, $\mathbf{u} \in \mathbb{R}^\ell$, $U \in \mathbb{R}$, define:

$$\hat{\phi}(U, \mathbf{A}) = \sum_{i=1}^{\ell} \frac{1}{U - \lambda_i},$$

and let $U(\mathbf{u}, \delta_U, \mathbf{A}, U)$ be defined as

$$U(\mathbf{u}, \delta_U, \mathbf{A}, U) = \frac{\mathbf{u}^\top (\mathbf{A} - U' \mathbf{I}_\ell)^{-2} \mathbf{u}}{\hat{\phi}(U, \mathbf{A}) - \hat{\phi}(U', \mathbf{A})} - \mathbf{u}^\top (\mathbf{A} - U' \mathbf{I}_\ell)^{-1} \mathbf{u},$$

where

$$U' = U + \delta_U = U + (1 + \ell/r) \left(1 - \sqrt{\ell/r}\right)^{-1}.$$

The running time of the algorithm is dominated by the search for an index i_τ satisfying

$$U(\mathbf{u}_{i_\tau}, \delta_U, \mathbf{A}_\tau, U_\tau) \leq \frac{1}{t_\tau} \leq L(\mathbf{u}_{i_\tau}, \delta_L, \mathbf{A}_\tau, L_\tau)$$

(one can achieve that by exhaustive search). One needs $\phi(L, \mathbf{A})$ and $\hat{\phi}(L, \mathbf{A})$, and hence the eigenvalues of \mathbf{A} . This takes $O(\ell^3)$ time, once per iteration, for a total of $O(r\ell^3)$. Then, for $i = 1, \dots, n$, we need to compute the functions L and U for every \mathbf{u}_i . This takes $O(n\ell^2)$ per iteration, for a total of $O(rn\ell^2)$. So, the total running time of the algorithm is $O(nr\ell^2)$.

3 Constrained Multiple-Response Regression

Constrained multiple-response regression in the Frobenius norm can be reduced to simple regression. So, we can apply the results of the previous section to this setting.

3.1 Multi-Objective Regression

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, with $\omega \geq 1$. The objective of multi-objective regression is:

$$\min_{\mathbf{x} \in \mathcal{D}} \|\mathbf{A}[\mathbf{x}, \dots, \mathbf{x}] - \mathbf{B}\|_{\text{F}}^2, \quad (3)$$

where $[\mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$ contains ω copies of $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$. Let $\mathbf{b}_{avg} = \frac{1}{\omega} \mathbf{B} \mathbf{1}_\omega$ (here $\mathbf{1}_\omega \in \mathbb{R}^\omega$ is a vector of all ones and thus $\mathbf{b}_{avg} \in \mathbb{R}^n$ is the average of the columns in \mathbf{B}). Recall that $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and let $\mathbf{X} = [\mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$.

Lemma 3. For $\mathbf{X} = [\mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$, $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{\text{F}}^2 = \omega \|\mathbf{A}\mathbf{x} - \mathbf{b}_{avg}\|_2^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|_2^2$.

In the above, $\mathbf{B}^{(i)} \in \mathbb{R}^n$ denotes the i -th column of \mathbf{B} as a column vector. Note that the second term in Lemma 3 does not depend on \mathbf{x} and thus the generalized multi-objective regression can be reduced to simple regression on \mathbf{A} and \mathbf{b}_{avg} . Using Theorem 1, we can get a coresets: let $\tilde{\mathbf{x}}_{opt}$ minimize $\|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b}_{avg})\|_2$, where \mathbf{S} and \mathbf{D} are obtained via Theorem 1 applied to \mathbf{A} and \mathbf{b}_{avg} . If $\tilde{\mathbf{X}}_{opt} = [\tilde{\mathbf{x}}_{opt}, \dots, \tilde{\mathbf{x}}_{opt}]$, then, by Lemma 3, $\tilde{\mathbf{X}}_{opt}$ minimizes $\|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{\text{F}}$. Similarly, if \mathbf{x}_{opt} minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}_{avg}\|_2$ and $\mathbf{X}_{opt} = [\mathbf{x}_{opt}, \dots, \mathbf{x}_{opt}]$, then \mathbf{X}_{opt} minimizes $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{\text{F}}$. Theorem 4 states that $\tilde{\mathbf{X}}_{opt}$ approximates \mathbf{X}_{opt} .

Theorem 4. Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, we can construct matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ (for any $r > k + 1$) such that the matrix $\tilde{\mathbf{X}}_{opt} = [\tilde{\mathbf{x}}_{opt}, \dots, \tilde{\mathbf{x}}_{opt}]$ that minimizes $\|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_F$ over all matrices $\mathbf{X} = [\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}]$ with $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$ satisfies:

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 \leq \left(1 + O\left(\sqrt{k/r}\right)\right) \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_F^2.$$

The run time of the proposed algorithm is $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}_{avg}]}) + O(n\omega + rnk^2)$, where $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}_{avg}]})$ is the time needed to compute the left singular vectors of the matrix $[\mathbf{A}, \mathbf{b}_{avg}] \in \mathbb{R}^{n \times (d+1)}$.

Proof. We first construct \mathbf{D} and \mathbf{S} via Theorem 1 applied to \mathbf{A} and \mathbf{b}_{avg} . The running time is $O(n\omega)$ (the time needed to compute \mathbf{b}_{avg}) plus the running time of Theorem 1. The result is immediate from the following derivation:

$$\begin{aligned} \|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 &\stackrel{(a)}{=} \omega \|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \\ &\stackrel{(b)}{\leq} \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \omega \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \\ &\leq \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \left(\omega \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \right) \\ &\stackrel{(a)}{=} \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_F^2. \end{aligned}$$

(a) follows by Lemma 3; (b) follows because $\tilde{\mathbf{x}}_{opt}$ is the output of a coresets regression as in Theorem 1. Finally, $r > k + 1$ implies that $\left(1 + O\left(\sqrt{k/r}\right)\right)^2 = 1 + O\left(\sqrt{k/r}\right)$. \blacksquare

3.2 Arbitrarily-Constrained Multiple-Response Regression

Multi-objective regression is a special case of constrained multiple-response regression for which we can efficiently obtain the coresets. In the general case, the problem still reduces to simple regression, but the coresets are now larger. The objective of arbitrarily-constrained multiple-response regression is

$$\min_{\mathbf{X} \in \mathcal{D} \subseteq \mathbb{R}^{d \times \omega}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F. \quad (4)$$

Since $\mathbb{R}^{d \times \omega}$ is isomorphic to $\mathbb{R}^{d\omega}$, we can view $\mathbf{X} \in \mathbb{R}^{d \times \omega}$ as a “stretched out” vector $\hat{\mathbf{X}} \in \mathbb{R}^{d\omega}$; corresponding to the domain \mathcal{D} is the domain $\hat{\mathcal{D}} \subseteq \mathbb{R}^{d\omega}$. Similarly, we can stretch out $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ to $\hat{\mathbf{B}} \in \mathbb{R}^{n\omega}$. To complete the transformation to simple linear regression, we build a transformed block-diagonal data matrix $\hat{\mathbf{A}}$ from \mathbf{A} , by repeating ω copies of \mathbf{A} along the diagonal:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & & & \\ & \mathbf{A} & & \\ & & \ddots & \\ & & & \mathbf{A} \end{bmatrix} \in \mathbb{R}^{n\omega \times d\omega}, \quad \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \\ \mathbf{X}^{(\omega)} \end{bmatrix} \in \mathbb{R}^{d\omega}, \quad \hat{\mathbf{B}} = \begin{bmatrix} \mathbf{B}^{(1)} \\ \mathbf{B}^{(2)} \\ \vdots \\ \mathbf{B}^{(\omega)} \end{bmatrix} \in \mathbb{R}^{n\omega}.$$

Lemma 5. For all \mathbf{A} , \mathbf{X} and \mathbf{B} of appropriate dimensions, $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 = \|\hat{\mathbf{A}}\hat{\mathbf{X}} - \hat{\mathbf{B}}\|_2^2$.

Theorem 1 gives us coresets for this equivalent regression. Note that $\text{rank}(\hat{\mathbf{A}}) \leq \omega \cdot \text{rank}(\mathbf{A})$. The coreset will identify the important rows of \mathbf{A} (the same row may get identified multiple times as different rows of

$\hat{\mathbf{A}}$), and the important *elements* of \mathbf{B} , because the entries in $\hat{\mathbf{B}}$ are elements of \mathbf{B} , not rows of \mathbf{B} . Let $\tilde{\mathbf{X}}_{opt}$ be the solution constructed from the coreset, which minimizes $\|\hat{\mathbf{A}}\tilde{\mathbf{X}} - \hat{\mathbf{B}}\|$ over $\tilde{\mathbf{X}} \in \hat{\mathcal{D}}$, and let $\tilde{\mathbf{X}}_{opt} \in \mathcal{D}$ be the corresponding solution in the original domain \mathcal{D} . If r is the size of the coreset and $\text{rank}(\mathbf{A}) = k$, then, by Theorem 1,

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\text{F}}^2 \leq \left(1 + O\left(\sqrt{k\omega/r}\right)\right) \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\text{F}}^2. \quad (5)$$

So, for the approximation ratio to be $1 + O(\epsilon)$, we set $r = O(k\omega/\epsilon^2)$. The running time would involve the time needed to compute the SVD of $[\hat{\mathbf{A}}, \hat{\mathbf{B}}]$.

Notice that the coresets are large and somewhat costly to compute and they only work for the Frobenius norm. In the next section, using more sophisticated techniques, we will get smaller coresets for unconstrained regression in both the Frobenius and spectral norms.

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$.
Output: sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.
 1: Compute the SVD of \mathbf{A} : $\mathbf{A} = \mathbf{U}_\mathbf{A} \Sigma_\mathbf{A} \mathbf{V}_\mathbf{A}^\text{T}$, where $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, $\Sigma_\mathbf{A} \in \mathbb{R}^{k \times k}$, and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d \times k}$; compute $\mathbf{E} = \mathbf{U}_\mathbf{A} \mathbf{U}_\mathbf{A}^\text{T} \mathbf{B} - \mathbf{B}$.
 2: **return** $[\mathbf{S}, \mathbf{D}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$ (see Lemma 10)

Algorithm 3: Deterministic coresets for multiple regression in spectral norm.

4 Unconstrained Multiple-Response Regression

Consider the following problem: given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank k and a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ with $\omega \geq 1$, we seek to identify the matrix $\mathbf{X}_{opt} \in \mathbb{R}^{d \times \omega}$ that satisfies ($\xi = 2$ and $\xi = \text{F}$)

$$\mathbf{X}_{opt} \in \arg \min_{\mathbf{X} \in \mathbb{R}^{d \times \omega}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{\xi}^2. \quad (6)$$

We can compute \mathbf{X}_{opt} via the pseudoinverse of \mathbf{A} , namely $\mathbf{X}_{opt} = \mathbf{A}^\dagger \mathbf{B}$. If \mathbf{S} and \mathbf{D} are sampling and rescaling matrices respectively, then the coreset regression problem is:

$$\tilde{\mathbf{X}}_{opt} \in \arg \min_{\mathbf{X} \in \mathbb{R}^{d \times \omega}} \|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{\xi}^2 = \arg \min_{\mathbf{X} \in \mathbb{R}^{d \times \omega}} \|\mathbf{D}\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{D}\mathbf{S}\mathbf{B}\|_{\xi}^2. \quad (7)$$

The solution of the coreset regression problem is $\tilde{\mathbf{X}}_{opt} = (\mathbf{D}\mathbf{S}\mathbf{A})^\dagger \mathbf{D}\mathbf{S}\mathbf{B}$. The main results in this section are presented in Theorems 6 and 7.

Theorem 6 (Spectral norm). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank k , a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$, Algorithm 3 deterministically constructs matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ such that the solution of the problem of Eqn. (7) satisfies:*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2 + \left(\frac{1 + \sqrt{\omega/r}}{1 - \sqrt{k/r}}\right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_\mathbf{A}) + O(rn(k^2 + \omega^2))$, where $T(\mathbf{U}_\mathbf{A})$ is the time needed to compute the left singular vectors of \mathbf{A} .

Since $r > k$, the approximation ratio is $2 + O(\sqrt{\omega/r} + \omega/r + \sqrt{k/r})$. So, for $\epsilon > 0$ and $r = O((\omega + k)/\epsilon^2)$ the approximation ratio is $2 + \epsilon$. For $r > \omega$, the approximation is $O(1)$, while for $r < \omega$, is asymptotic to $O(\omega/r)$. We will argue that this is nearly optimal by providing a matching lower bound in Theorem 15.

Theorem 7 (Frobenius norm). *Given matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$, Algorithm 4 deterministically constructs a sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ and a rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$ such that the solution of the problem of Eqn. (7) satisfies:*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\text{F}}^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\text{F}}^2 + \frac{1}{(1 - \sqrt{k/r})^2} \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\text{F}}^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_{\mathbf{A}}) + O(rnk^2)$, where $T(\mathbf{U}_{\mathbf{A}})$ is the time needed to compute the left singular vectors of \mathbf{A} .

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$.
Output: sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.
 1: Compute the SVD of \mathbf{A} : $\mathbf{A} = \mathbf{U}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^{\text{T}}$, where $\mathbf{U}_{\mathbf{A}} \in \mathbb{R}^{n \times k}$, $\mathbf{\Sigma}_{\mathbf{A}} \in \mathbb{R}^{k \times k}$, and $\mathbf{V}_{\mathbf{A}} \in \mathbb{R}^{d \times k}$; compute $\mathbf{E} = \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^{\text{T}}\mathbf{B} - \mathbf{B}$.
 2: **return** $[\mathbf{S}, \mathbf{D}] = \text{MultipleFrobeniusSampling}(\mathbf{U}_{\mathbf{A}}, \mathbf{E}, r)$ (see Lemma 11)

Algorithm 4: Deterministic coresets for multiple regression in Frobenius norm.

The approximation ratio in the above theorem is $2 + O(\sqrt{k/r})$. In Theorem 16, we will give a lower bound for the approximation ratio which is $1 + \Omega(k/r)$. We conjecture that our lower bound can be achieved (deterministically), perhaps by a more sophisticated algorithm or analysis.

Finally, we note that the **B-agnostic** randomized construction of [10] achieves a $(1 + \epsilon)$ approximation ratio using a significantly larger coreset, $r = O(k \log k / \epsilon^2)$. Importantly, [10] does not need any access to \mathbf{B} in order to construct the coreset, whereas our approach constructs coresets by carefully choosing important data points with respect to the particular target response matrix \mathbf{B} . We will also discuss **B-agnostic** algorithms in Section 4.2 (Theorem 12) and we will present matching lower bounds in Section 5.

4.1 Proofs of Theorems 6 and 7

We will make heavy use of facts from Section A in the Appendix. We start with a few simple lemmas.

Lemma 8. *Let $\mathbf{E} = \mathbf{A}\mathbf{X}_{opt} - \mathbf{B} \in \mathbb{R}^{n \times \omega}$ be the regression residual. Then, $\text{rank}(\mathbf{E}) \leq \min\{\omega, n - k\}$.*

Proof. Using our notation, $\mathbf{A}\mathbf{X}_{opt} - \mathbf{B} = -(\mathbf{I}_n - \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^{\text{T}})\mathbf{B} = -\mathbf{U}_{\mathbf{A}}^{\perp}(\mathbf{U}_{\mathbf{A}}^{\perp})^{\text{T}}\mathbf{B}$. To conclude notice that $\text{rank}(\mathbf{X}\mathbf{Y}) \leq \min\{\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y})\}$ for any matrices \mathbf{X} and \mathbf{Y} . ■

We now present our main tool for obtaining approximation guarantees for coreset regression.

Lemma 9. *Assume that the rank of the matrix $\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{A}} \in \mathbb{R}^{r \times k}$ is equal to k (i.e., the matrix has full rank). Then, for $\xi = 2, \text{F}$,*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\xi}^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\xi}^2 + \|(\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{A}})^{\dagger}\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\xi}^2.$$

Proof. To simplify notation, let $\mathbf{W} = \mathbf{D}\mathbf{S}$. Using the SVD of \mathbf{A} , $\mathbf{A} = \mathbf{U}_\mathbf{A}\boldsymbol{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top$, we get:

$$\|\mathbf{B} - \mathbf{A}\tilde{\mathbf{X}}_{opt}\|_\xi^2 = \|\mathbf{B} - \mathbf{U}_\mathbf{A}\boldsymbol{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{W}\mathbf{U}_\mathbf{A}\boldsymbol{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top)^\dagger\mathbf{W}\mathbf{B}\|_\xi^2 = \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{B}\|_\xi^2,$$

where the last equality follows from properties of the pseudo-inverse and the fact that $\mathbf{W}\mathbf{U}_\mathbf{A}$ is a full-rank matrix (see Lemma 18 in the Appendix). Using $\mathbf{B} = \left(\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top + \mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\right)\mathbf{B}$, we obtain

$$\begin{aligned} \|\mathbf{B} - \mathbf{A}\tilde{\mathbf{X}}_{opt}\|_\xi^2 &= \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\left(\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top + \mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\right)\mathbf{B}\|_\xi^2 \\ &= \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top\mathbf{B} + \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B}\|_\xi^2 \\ &\stackrel{(a)}{=} \|\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B} + \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B}\|_\xi^2 \\ &\stackrel{(b)}{\leq} \|\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B}\|_\xi^2 + \|\mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B}\|_\xi^2. \end{aligned}$$

(a) follows from the assumption that the rank of $\mathbf{W}\mathbf{U}_\mathbf{A}$ is equal to k and thus $(\mathbf{W}\mathbf{U}_\mathbf{A})^\dagger\mathbf{W}\mathbf{U}_\mathbf{A} = \mathbf{I}_k$ and (b) follows by matrix-Pythagoras (Lemma 17). To conclude, we use spectral submultiplicativity on the second term and the fact that $\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\top\mathbf{B} = -(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})$. \blacksquare

This lemma provides a framework for coresets construction: all we need are sampling and rescaling matrices \mathbf{S} and \mathbf{D} , such that $\text{rank}(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A}) = k$ and

$$\|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_\xi^2$$

is small. The final ingredients for the proofs of Theorems 6 and 7 are two matrix sparsification results that we present in the Appendix.

Lemma 10. *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_\mathbf{Y}$, and $\rho_\boldsymbol{\Psi}$. Given $r > \rho_\mathbf{Y}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + T_{\text{SVD}}(\boldsymbol{\Psi}) + O(rn(\rho_\mathbf{Y}^2 + \rho_\boldsymbol{\Psi}^2))$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{D}\mathbf{S}\mathbf{Y}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{D}\mathbf{S}\mathbf{Y})^\dagger\|_2 < \frac{1}{1 - \sqrt{\rho_\mathbf{Y}/r}}\|\mathbf{Y}^\dagger\|_2; \quad \|\mathbf{D}\mathbf{S}\boldsymbol{\Psi}\|_2 < \left(1 + \sqrt{\frac{\rho_\boldsymbol{\Psi}}{r}}\right)\|\boldsymbol{\Psi}\|_2.$$

If $\boldsymbol{\Psi} = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{Y}, \boldsymbol{\Psi}, r)$ to denote such a deterministic procedure.

Lemma 11. *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_\mathbf{Y}$, and $\rho_\boldsymbol{\Psi}$. Given $r > \rho_\mathbf{Y}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2 + \ell_2 n)$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{D}\mathbf{S}\mathbf{Y}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{D}\mathbf{S}\mathbf{Y})^\dagger\|_2 < \frac{1}{1 - \sqrt{\rho_\mathbf{Y}/r}}\|\mathbf{Y}^\dagger\|_2; \quad \|\mathbf{D}\mathbf{S}\boldsymbol{\Psi}\|_F \leq \|\boldsymbol{\Psi}\|_F.$$

If $\boldsymbol{\Psi} = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleFrobeniusSampling}(\mathbf{Y}, \boldsymbol{\Psi}, r)$ to denote such a deterministic procedure.

Proof. (of Theorem 6) Theorem 6 follows from Lemmas 9 and 10. First, compute the SVD of \mathbf{A} to obtain $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, and let $\mathbf{E} = \mathbf{A}\mathbf{X}_{opt} - \mathbf{B} = \mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top\mathbf{B} - \mathbf{B}$. Next, run the algorithm of Lemma 10 to obtain $[\mathbf{D}, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$. This algorithm runs in time $T_{SVD}(\mathbf{E}) + O(rn(k^2 + \rho_{\mathbf{E}}^2))$, where k is the rank of $\mathbf{U}_\mathbf{A}$ and \mathbf{A} . The total running time of the algorithm is $T(\mathbf{U}_\mathbf{A}) + T_{SVD}(\mathbf{E}) + O(rn(k^2 + \rho_{\mathbf{E}}^2)) = T(\mathbf{U}_\mathbf{A}) + O(rn(k^2 + \omega^2))$.

Lemma 10 guarantees that \mathbf{D} and \mathbf{S} satisfy the rank assumption of Lemma 9. To conclude the proof, we bound the second term of Lemma 9, using the bounds of Lemma 10 and $\rho_{\mathbf{E}} \leq \min\{\omega, n - k\} \leq \omega$:

$$\begin{aligned} \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger \mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_2^2 &\leq \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger\|_2^2 \|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_2^2 \\ &\leq \left(1 - \sqrt{k/r}\right)^{-2} \left(1 + \sqrt{\omega/r}\right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2. \end{aligned}$$

■

Proof. (of Theorem 7) The proof is similar to the proof of Theorem 6, using Lemma 11 instead of Lemma 10. Let $[\mathbf{D}, \mathbf{S}] = \text{MultipleFrobeniusSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$. We bound the second term of Lemma 9, using the bounds of Lemma 11:

$$\begin{aligned} \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger \mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\mathbb{F}}^2 &\leq \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger\|_{\mathbb{F}}^2 \|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\mathbb{F}}^2 \\ &\leq \left(1 - \sqrt{k/r}\right)^{-2} \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\mathbb{F}}^2. \end{aligned}$$

■

4.2 B-Agnostic Coreset Construction

All the coreset construction algorithms that we presented so far carefully construct the coreset using knowledge of the response vector. If the algorithm does not need knowledge of \mathbf{B} to construct the coreset, and yet can provide an approximation guarantee for every \mathbf{B} , then the algorithm is \mathbf{B} -agnostic. A \mathbf{B} -agnostic coreset construction algorithm is appealing because the coreset, as specified by the sampling and rescaling matrices \mathbf{S} and \mathbf{D} , can be computed off-line and applied to any \mathbf{B} . We briefly digress to show how our methods can be extended to develop \mathbf{B} -agnostic coreset constructions.

Theorem 12 (**B-agnostic Coresets**). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank k , a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$, there exists an algorithm to deterministically construct a sampling matrix \mathbf{S} and a rescaling matrix \mathbf{D} such that for any $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, the matrix $\tilde{\mathbf{X}}_{opt}$ that solves the problem of Eqn. (7) satisfies:*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\xi}^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\xi}^2 + \left(\frac{1 + \sqrt{n/r}}{1 - \sqrt{k/r}}\right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\xi}^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_\mathbf{A}) + O(rnk^2)$, where $T(\mathbf{U}_\mathbf{A})$ is the time needed to compute the left singular vectors of \mathbf{A} .

Proof. The proof is similar to the proof of Theorem 6, except we now construct the sampling and rescaling matrices as $[\mathbf{S}, \mathbf{D}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{I}_n, r)$. To bound the second term in Lemma 9, we use

$$\begin{aligned} \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger \mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\xi}^2 &= \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger \mathbf{D}\mathbf{S}\mathbf{I}_n(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\xi}^2 \\ &\leq \|(\mathbf{D}\mathbf{S}\mathbf{U}_\mathbf{A})^\dagger\|_2^2 \|\mathbf{D}\mathbf{S}\mathbf{I}_n\|_2^2 \|(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_{\xi}^2, \end{aligned}$$

and the bounds of Lemma 10. ■

The above bound decreases with r and holds for any \mathbf{B} , guaranteeing a constant-factor approximation with a constant fraction of the data. The approximation ratio is $O(n/r)$, which seems quite weak. In the next section, we show that this result is indeed tight.

Type of Regression	Lower bound	Known Approximation Ratio
Deterministic \mathbf{b} -agnostic	n/r [Thm. 13]	$O(n/r)$ [Thm. 12]
Randomized \mathbf{b} -agnostic	$1 + \Omega(1/r)$ [Thm. 14]	$1 + O(\sqrt{k \log k/r})$ [Thm. 5 in [10]]
Multiple-regression ($\xi = 2$)	$\omega/(r + 1)$ [Thm. 15]	$2 + O(\sqrt{\omega/r} + \omega/r + \sqrt{k/r})$ [Thm. 6]
Multiple-regression ($\xi = F$)	$1 + \Omega(k/r)$ [Thm. 16]	$2 + O(\sqrt{k/r})$ [Thm. 7]

Table 2: Lower bounds on the approximation ratio for different formulations of linear regression and a coreset of size r . The randomized algorithm in the second row of the table delivers a constant probability of success (all other algorithms are deterministic). The lower bounds are values γ such that $\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|/\|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\| \geq \gamma$. The approximation ratios are values β such that $\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|/\|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\| \leq \beta$. In the first two rows in the table, \mathbf{X}_{opt} , $\tilde{\mathbf{X}}_{opt}$, and \mathbf{B} are vectors. NOTATION: n is the number of data points of dimension $d < n$; k is the rank of the matrix whose rows correspond to the n data points; r is the size of the coreset, $k < r < n$; $\omega \geq 1$ is the number of “response” vectors in multiple-response regression (in the last two rows in the table \mathbf{X}_{opt} , $\tilde{\mathbf{X}}_{opt}$, and \mathbf{B} have ω columns).

5 Lower Bounds on Coreset Size

We have just seen a \mathbf{B} -agnostic coreset construction algorithm with a rather weak worst case guarantee of $O(n/r)$ approximation error. We will now show that no deterministic \mathbf{B} -agnostic coreset construction algorithm can guarantee a better error (Theorem 13) by providing lower bounds on coreset size as a function of approximation error. These results are also summarized in Table 2.

[10] provides another \mathbf{B} -agnostic coreset construction algorithm with $r = O(k \log k/\epsilon^2)$. For a fixed \mathbf{B} , the method in [10] delivers a probabilistic bound on the approximation error. However, there are target matrices \mathbf{B} for which the bound fails by an arbitrarily large amount. The probabilistic algorithms get away with this by brushing all these (possibly large) errors into a low probability event, with respect to random choices made in the algorithm. So, in some sense, these algorithms are not \mathbf{B} -agnostic, in that they do not construct a coreset which works well for all \mathbf{B} with some (say) constant probability. Nevertheless, the fact that they give a constant probability of success for a fixed but unknown \mathbf{B} makes these algorithms interesting and useful. We will give a lower bound on the approximation ratio of such algorithms as well, for a given probability of success (Theorem 14). Finally, we will give lower bounds on the size of the coreset for the general (non-agnostic) multiple regression setting (Theorems 15 and 16).

5.1 An Impossibility Result for \mathbf{B} -Agnostic Coreset Construction

We first present the lower bound for simple regression. Recall that a coreset construction algorithm is \mathbf{b} -agnostic if it constructs a coreset without knowledge of \mathbf{b} , and then provides an approximation guarantee for every \mathbf{b} . We show that no coreset can work for every \mathbf{b} ; therefore a \mathbf{b} -agnostic coreset will be bad for some vector \mathbf{b} . In fact, there exists a matrix \mathbf{A} such that every coreset has an associated “bad” \mathbf{b} .

Theorem 13 (Deterministic \mathbf{b} -Agnostic coresets). *There exists a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ such that for every coreset $\mathbf{C} \in \mathbb{R}^{r \times d}$ of size $r \leq n$, there exists $\mathbf{b} \in \mathbb{R}^n$ (depending on \mathbf{C}) for which*

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{r} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

Proof. Let \mathbf{A} be any matrix with orthonormal columns whose first column is $\mathbf{1}_n/\sqrt{n}$, and consider any coreset \mathbf{C} of size r . Let $\mathbf{b} = \mathbf{1}_{\overline{\mathbf{C}}}/\sqrt{n-r}$, where $\mathbf{1}_{\overline{\mathbf{C}}}$ is the n -vector of 1's except at the coreset locations. So for the coreset regression, $\mathbf{b}_c = \mathbf{0}$, and so $\tilde{\mathbf{x}}_{opt} = \mathbf{0}_{d \times 1}$. Therefore,

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 = \|\mathbf{b}\|_2^2 = 1.$$

Let $\mathbf{P}_{\mathbf{A}}$ project onto the columns of \mathbf{A} and $\mathbf{P}_{\mathbf{A}^{(1)}}$ project onto the first column of \mathbf{A} . The following sequence establishes the result:

$$\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{b}\|_2^2 \leq \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{(1)}})\mathbf{b}\|_2^2 = \frac{r}{n}$$

■

We now consider randomized algorithms that construct a coreset without looking at \mathbf{b} (e.g. [10]). These algorithms work for any fixed (but unknown) \mathbf{b} , and deliver a probabilistic approximation guarantee for any single fixed \mathbf{b} ; in some sense they are \mathbf{b} -agnostic. By the previous discussion, the returned coreset must fail for some \mathbf{b} , i.e., the probabilistic guarantee does not hold for all \mathbf{b} , and, when it fails, it could do so with very bad error. We will now present a lower bound on the approximation accuracy of such existing randomized algorithms for coreset construction, even for a single \mathbf{b} .

First, we define randomized coreset construction algorithms. Let $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_{\binom{n}{r}}$ be the $\binom{n}{r}$ different coresets of size r . A randomized algorithm assigns probabilities $p_1, p_2, \dots, p_{\binom{n}{r}}$ to each coreset, and selects one according to these probabilities. The probabilities p_i may depend on \mathbf{A} . The algorithm is \mathbf{b} -agnostic if the probabilities p_i do not depend on \mathbf{b} . As usual, let r be the size of the coreset.

Theorem 14 (Probabilistic \mathbf{b} -Agnostic Coresets). *For any randomized \mathbf{b} -agnostic coreset construction algorithm, and any integer $0 \leq \ell \leq n - r$, there exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$, such that, with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$,*

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{n-\ell} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

Proof. Let \mathbf{A} be any matrix with orthonormal columns whose first column is $\mathbf{1}_n/\sqrt{n}$, as in the proof of Theorem 13. Let \mathbf{T} be a set of size $\ell \leq n - r$. The neighborhood $N(\mathbf{T})$ is the set of coresets (of size r) that have non-empty intersection with \mathbf{T} . Every coreset appears in $\binom{n}{\ell} - \binom{n-r}{\ell}$ such neighborhoods (the number of sets of size ℓ which intersect with a coreset of size r). Let \mathbf{C} be the random coreset (of size r) selected by the algorithm. Let $\Pr[\mathbf{C} \in N(\mathbf{T})]$ be the probability that the coreset selected by the algorithm is in $N(\mathbf{T})$; then, $\Pr[\mathbf{C} \in N(\mathbf{T})] = \sum_{\mathbf{C}_i \in N(\mathbf{T})} \Pr[\mathbf{C}_i]$. Therefore,

$$\sum_{\mathbf{T}} \Pr[\mathbf{C} \in N(\mathbf{T})] = \sum_{\mathbf{T}} \sum_{\mathbf{C}_i \in N(\mathbf{T})} \Pr[\mathbf{C}_i] = \binom{n}{\ell} - \binom{n-r}{\ell},$$

where the last equality follows because each coreset appears exactly $\binom{n}{\ell} - \binom{n-r}{\ell}$ times in the summation and $\sum_i \Pr[\mathbf{C}_i] = 1$. Thus, there is at least one set \mathbf{T}^* for which

$$\Pr[\mathbf{C} \in N(\mathbf{T}^*)] \leq \frac{\binom{n}{\ell} - \binom{n-r}{\ell}}{\binom{n}{\ell}} = 1 - \frac{\binom{n-r}{\ell}}{\binom{n}{\ell}}.$$

So, with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$, the selected coreset does not intersect with \mathbf{T}^* . Select $\mathbf{b} = \mathbf{1}_{\mathbf{T}^*}$ (the unit vector which is $1/\sqrt{\ell}$ at the indices corresponding to \mathbf{T}^*). Now, with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$, $\tilde{\mathbf{x}}_{opt} = \mathbf{0}$, and the analysis in the proof of Theorem 13 shows that

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{n-\ell} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

■

By Stirling's formula, after some algebra, the probability $\binom{n-r}{\ell} / \binom{n}{\ell}$ is asymptotic to $e^{-2r\ell/n}$. Setting $\ell = \Theta(n/r)$ gives a success probability that is a constant. Then, the approximation ratio cannot be better than $1 + \Omega(1/r)$. With regard to high probability (approaching one) algorithms, consider $\ell = n \log n / 2r$ to conclude that if the success probability is at least $1 - 1/n$, the approximation ratio is no better than $1 + \log(n)/(2r - \log n)$.

5.2 Lower Bounds for Non-Agnostic Multiple Regression

For both the spectral and the Frobenius norm, we now consider non-agnostic unconstrained multiple regression, and give lower bounds for coresets of size $r > d = \text{rank}(\mathbf{A})$ (for simplicity, we set $\text{rank}(\mathbf{A}) = d$). The results are presented in Theorems 15 and 16.

Theorem 15 (Spectral Norm). *There exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank d and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ such that for any $r > d$ and any sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$, the solution to the coreset regression $\tilde{\mathbf{X}}_{opt} = (\mathbf{D}\mathbf{S}\mathbf{A})^\dagger \mathbf{D}\mathbf{S}\mathbf{B} \in \mathbb{R}^{d \times \omega}$ satisfies*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 \geq \frac{\omega}{r+1} \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2.$$

Proof. First, we need some results from [7]. Consider the matrix

$$\mathbf{H} = [\mathbf{e}_1 + \alpha\mathbf{e}_2, \mathbf{e}_1 + \alpha\mathbf{e}_3, \dots, \mathbf{e}_1 + \alpha\mathbf{e}_\omega] \in \mathbb{R}^{\omega \times (\omega-1)},$$

where $\mathbf{e}_i \in \mathbb{R}^\omega$ are the standard basis vectors. Then, let $\mathbf{B} = \mathbf{H}^\top \in \mathbb{R}^{(\omega-1) \times \omega}$. Theorem 34 in [7] (with $\alpha = 1$) argues the following: given \mathbf{B} and any sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times (\omega-1)}$ and diagonal rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$, with $\hat{\mathbf{C}} = \mathbf{D}\mathbf{S}\mathbf{B}$ (rescaled sampled coreset of \mathbf{B}), and any k with $1 \leq k \leq \omega - 1$,

$$\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_2^2 \geq \frac{\omega}{r+1} \|\mathbf{B} - \mathbf{B}_k\|_2^2.$$

In the above, $\Pi_{\hat{\mathbf{C}},k}(\mathbf{B}) \in \mathbb{R}^{(\omega-1) \times \omega}$ of rank k is the best rank- k approximation to \mathbf{B} (in the spectral norm) whose rows lie in the span of all the rows in $\hat{\mathbf{C}}$ (the row-space of $\hat{\mathbf{C}}$); and, $\mathbf{B}_k \in \mathbb{R}^{(\omega-1) \times \omega}$ of rank k is the best rank- k approximation to \mathbf{B} (which could be computed via the truncated SVD of \mathbf{B}).¹

Since $\Pi_{\hat{\mathbf{C}},k}(\mathbf{B})$ is the best rank- k approximation to \mathbf{B} in the row-space of $\hat{\mathbf{C}}$, it follows that

$$\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_2^2 \leq \|\mathbf{B} - \mathbf{X}\hat{\mathbf{C}}\|_2^2,$$

for any $\mathbf{X} \in \mathbb{R}^{(\omega-1) \times r}$ with rank at most k (because $\mathbf{X}\hat{\mathbf{C}}$ will have rank at most k and is in the row space of $\hat{\mathbf{C}}$). Set $\mathbf{X} = \mathbf{U}_{\mathbf{B},k}(\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{B},k})^\dagger$, where $\mathbf{U}_{\mathbf{B},k} \in \mathbb{R}^{(\omega-1) \times k}$ has k columns which are the top- k left singular vectors of \mathbf{B} . It is easy to verify that \mathbf{X} has the correct dimensions and rank at most k . Since $\hat{\mathbf{C}} = \mathbf{D}\mathbf{S}\mathbf{B}$, we have that

$$\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_2^2 \leq \|\mathbf{B} - \mathbf{U}_{\mathbf{B},k}(\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{B},k})^\dagger \mathbf{D}\mathbf{S}\mathbf{B}\|_2^2.$$

We now construct the regression problem which exhibits the lower bound in the theorem. Let $\mathbf{A} = \mathbf{U}_{\mathbf{B},d} \in \mathbb{R}^{(\omega-1) \times d}$ (i.e., we choose $k = d$ in the above discussion) and $n = \omega - 1$. \mathbf{B} is as we described above. Suppose a coreset construction algorithm gives sampling and rescaling matrices \mathbf{S} and \mathbf{D} , for a coreset of size r . So, the coreset regression is with $\tilde{\mathbf{A}} = \mathbf{D}\mathbf{S}\mathbf{A} = \mathbf{C}$ and $\tilde{\mathbf{B}} = \mathbf{D}\mathbf{S}\mathbf{B}$. The solution to the coreset regression is

$$\tilde{\mathbf{X}}_{opt} = \mathbf{C}^\dagger \mathbf{D}\mathbf{S}\mathbf{B} = (\mathbf{D}\mathbf{S}\mathbf{A})^\dagger \mathbf{D}\mathbf{S}\mathbf{B} = (\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{B},d})^\dagger \mathbf{D}\mathbf{S}\mathbf{B},$$

¹Actually, \mathbf{D} is irrelevant here because the row-space of $\mathbf{S}\mathbf{B}$ is the same as the row space of $\mathbf{D}\mathbf{S}\mathbf{B}$.

which means that

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 = \|\mathbf{U}_{\mathbf{B},d}(\mathbf{DSU}_{\mathbf{B},d})^\dagger \mathbf{DSB} - \mathbf{B}\|_2^2 \geq \|\Pi_{\mathbf{C},d}(\mathbf{B}) - \mathbf{B}\|_2^2 \geq \frac{\omega}{r+1} \|\mathbf{B}_d - \mathbf{B}\|_2^2.$$

To conclude the proof, observe that $\mathbf{B}_d = \mathbf{U}_{\mathbf{B},d} \mathbf{U}_{\mathbf{B},d}^\top \mathbf{B} = \mathbf{A} \mathbf{A}^\dagger \mathbf{B} = \mathbf{A} \mathbf{X}_{opt}$. \blacksquare

Theorem 16 (Frobenius Norm). *There exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank d and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ such that for any $r > d$ and any sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{n \times r}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$, the solution to the coreset regression $\tilde{\mathbf{X}}_{opt} = (\mathbf{DSA})^\dagger \mathbf{DSB} \in \mathbb{R}^{d \times \omega}$ satisfies (for any $\alpha > 0$)*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\mathbb{F}}^2 \geq \frac{n-r}{n-d} \left(1 + \frac{d}{r + \alpha^2}\right) \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\mathbb{F}}^2.$$

As $\alpha \rightarrow 0$ and $n \rightarrow \infty$ the lower bound is $1 + d/r$.

Proof. First, we need some results from [7]. For any integer $\gamma > 1$ and any integer $k \geq 1$, Theorem 36 in [7] exhibits a matrix $\mathbf{B} \in \mathbb{R}^{\gamma k \times (\gamma+1)k}$ such that for any sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times \gamma k}$ and diagonal rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$, with $\hat{\mathbf{C}} = \mathbf{DSB}$ (rescaled sampled coreset of \mathbf{B}), any $\alpha > 0$, and any $r \geq 1$,

$$\frac{\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_{\mathbb{F}}^2}{\|\mathbf{B} - \mathbf{B}_k\|_{\mathbb{F}}^2} \geq \frac{\gamma k - r}{\gamma k - k} \left(1 + \frac{k}{r + \alpha^2}\right).$$

The matrix \mathbf{B} is constructed as follows. Recall that γ is any positive integer with $\gamma > 1$. Let \mathbf{A} have dimensions $(\gamma + 1) \times \gamma$ and be constructed as follows.

$$\mathbf{A} = \left[\mathbf{e}_1 + \frac{\alpha}{\sqrt{k}} \mathbf{e}_2, \mathbf{e}_1 + \frac{\alpha}{\sqrt{k}} \mathbf{e}_3, \dots, \mathbf{e}_1 + \frac{\alpha}{\sqrt{k}} \mathbf{e}_\gamma \right],$$

where $\mathbf{e}_i \in \mathbb{R}^{\gamma+1}$ are the standard basis vectors. Now construct \mathbf{H} to be block diagonal, with k copies of \mathbf{A} along its diagonal; so, the dimensions of \mathbf{H} are $(\gamma + 1)k \times \gamma k$. Then, $\mathbf{B} = \mathbf{H}^\top$.

In the above, $\Pi_{\hat{\mathbf{C}},k}(\mathbf{B}) \in \mathbb{R}^{\gamma k \times (\gamma+1)k}$ of rank k is the best rank- k approximation to \mathbf{B} (in the Frobenius norm) whose rows lie in the span of all the rows in $\hat{\mathbf{C}}$ (the row-space of $\hat{\mathbf{C}}$); and, $\mathbf{B}_k \in \mathbb{R}^{\gamma k \times (\gamma+1)k}$ of rank k is the best rank- k approximation to \mathbf{B} (which could be computed via the truncated SVD of \mathbf{B}). Since $\Pi_{\hat{\mathbf{C}},k}(\mathbf{B})$ is the best rank- k approximation to \mathbf{B} in the row-space of $\hat{\mathbf{C}}$, it follows that

$$\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_{\mathbb{F}}^2 \leq \|\mathbf{B} - \mathbf{X}\hat{\mathbf{C}}\|_{\mathbb{F}}^2,$$

for any $\mathbf{X} \in \mathbb{R}^{\gamma k \times r}$ with rank at most k (because $\mathbf{X}\hat{\mathbf{C}}$ will have rank at most k and is in the row space of $\hat{\mathbf{C}}$). Set $\mathbf{X} = \mathbf{U}_{\mathbf{B},k}(\mathbf{DSU}_{\mathbf{B},k})^\dagger$, where $\mathbf{U}_{\mathbf{B},k} \in \mathbb{R}^{\gamma k \times k}$ has k columns which are the top- k left singular vectors of \mathbf{B} . It is easy to verify that \mathbf{X} has the correct dimensions and rank at most k . Since $\hat{\mathbf{C}} = \mathbf{DSB}$, we have that

$$\|\mathbf{B} - \Pi_{\hat{\mathbf{C}},k}(\mathbf{B})\|_{\mathbb{F}}^2 \leq \|\mathbf{B} - \mathbf{U}_{\mathbf{B},k}(\mathbf{DSU}_{\mathbf{B},k})^\dagger \mathbf{DSB}\|_{\mathbb{F}}^2.$$

We now construct the regression problem which proves the lower bound in the theorem. Let $\mathbf{A} = \mathbf{U}_{\mathbf{B},d} \in \mathbb{R}^{\gamma d \times d}$ (i.e., we choose $k = d$ in the above discussion), $n = \gamma d$ (i.e. n is a multiple of d in the regression problem), and $\omega = (\gamma + 1)d$. \mathbf{B} is as we described above. Suppose a coreset construction algorithm gives sampling and rescaling matrices \mathbf{S} and \mathbf{D} , for a coreset of size $r > d$. So, the coreset regression is with $\mathbf{C} = \mathbf{DSA} \in \mathbb{R}^{r \times d}$ and $\mathbf{DSB} \in \mathbb{R}^{r \times \omega}$. The solution to the coreset regression is

$$\tilde{\mathbf{X}}_{opt} = \mathbf{C}^\dagger \mathbf{DSB} = (\mathbf{DSA})^\dagger \mathbf{DSB} = (\mathbf{DSU}_{\mathbf{B},d})^\dagger \mathbf{DSB},$$

which means that

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\mathbb{F}}^2 = \|\mathbf{U}_{\mathbf{B},d}(\mathbf{DSU}_{\mathbf{B},d})^\dagger \mathbf{DSB} - \mathbf{B}\|_{\mathbb{F}}^2 \geq \|\Pi_{\mathbf{C},d}(\mathbf{B}) - \mathbf{B}\|_{\mathbb{F}}^2 \geq \frac{\omega - d - r}{\omega - 2d} \left(1 + \frac{d}{r + \alpha^2}\right) \|\mathbf{B}_d - \mathbf{B}\|_{\mathbb{F}}^2.$$

To conclude the proof, observe that $\mathbf{B}_d = \mathbf{U}_{\mathbf{B},d} \mathbf{U}_{\mathbf{B},d}^\top \mathbf{B} = \mathbf{A} \mathbf{A}^\dagger \mathbf{B} = \mathbf{A} \mathbf{X}_{opt}$ and $\omega = n + d$. \blacksquare

6 Open problems

An important open problem arises in our work: can we determine the minimum size of a coresets that provides a $(1 + \epsilon)$ relative-error guarantee for simple linear regression? We conjecture that $\Omega(k/\epsilon)$ is a lower bound, which will make our results almost tight. Certainly, coresets of size exactly k cannot be guaranteed: consider two data points $(1, 1), (-1, 1)$. The optimal regression is zero; however any coreset of size one will give non-zero regression.

Acknowledgements. Christos Boutsidis acknowledges the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. Petros Drineas and Malik Magdon-Ismael have been supported by NSF CCF 1016501, NSF DMS 1008983, and NSF CCF CAREER 824684.

References

- [1] A. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- [2] J.D. Batson, D.A. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proc. 41st Annual ACM STOC*, pages 255–262, 2009.
- [3] S. Bellavia, M. Macconi, and B. Morini. An interior point newton-like method for non-negative least squares problems with degenerate solution. *Numerical Linear Algebra with Applications*, 13:825–844, 2006.
- [4] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas with Applications to Linear Systems Theory*. Princeton University Press, 2005.
- [5] C. Boutsidis. *Topics in Matrix Sampling Algorithms*. PhD thesis, Rensselaer Polytechnic Institute, 2011. <http://arxiv.org/abs/1105.0709>.
- [6] C. Boutsidis and P. Drineas. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431(5-7):760–771, 2009.
- [7] C. Boutsidis, P. Drineas, and M. Magdon-Ismael. Near-optimal column based matrix reconstruction. *Preprint, Available online, ArXiv*, 2011.
- [8] L. Breiman and J. Friedman. Predicting multivariate responses in multiple linear regression. *J. Royal Stat. Soc.*, 59(1):3–54, 1997.
- [9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.
- [10] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal Matrix Analysis and Applications*, 30(2):844–881, 2008.
- [11] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proc. SODA*, pages 1127–1136, 2006.
- [12] D.Y. Gao. Solutions and optimality criteria to box constrained nonconvex minimization problems. *MANAGEMENT*, 3(2):293–304, 2007.

- [13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [14] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [15] R. Kannan and S. Vempala. Nimble Algorithms for Cloud Computing. arXiv preprint arXiv:1304.3162, 2013.
- [16] C. L. Lawson and R. J. Hanson. Solving least squares problems. *Prentice-Hall*, 1974.
- [17] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. of the ACM*, 54, 2007.
- [18] G.A.F. Seber and A.J. Lee. *Linear regression analysis*. Wiley New York, 1977.

A Linear Algebra Background

The Singular Value Decomposition (SVD) of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k is a decomposition

$$\mathbf{A} = \mathbf{U}_\mathbf{A} \boldsymbol{\Sigma}_\mathbf{A} \mathbf{V}_\mathbf{A}^\top.$$

The singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > 0$ are contained in the diagonal matrix $\boldsymbol{\Sigma}_\mathbf{A} \in \mathbb{R}^{k \times k}$; $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$ contains the left singular vectors of \mathbf{A} ; and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d \times k}$ contains the right singular vectors. The Moore-Penrose pseudo-inverse of \mathbf{A} is $\mathbf{A}^\dagger = \mathbf{V}_\mathbf{A} \boldsymbol{\Sigma}_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A}^\top$. Given an orthonormal matrix $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, the perpendicular matrix $\mathbf{U}_\mathbf{A}^\perp \in \mathbb{R}^{n \times (n-k)}$ to $\mathbf{U}_\mathbf{A}$ satisfies: $(\mathbf{U}_\mathbf{A}^\perp)^\top \mathbf{U}_\mathbf{A}^\perp = \mathbf{I}_{n-k}$, $\mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{A}^\perp = \mathbf{0}_{k \times (n-k)}$, and $\mathbf{U}_\mathbf{A} \mathbf{U}_\mathbf{A}^\top + \mathbf{U}_\mathbf{A}^\perp (\mathbf{U}_\mathbf{A}^\perp)^\top = \mathbf{I}_n$. All the singular values of both $\mathbf{U}_\mathbf{A}$ and $\mathbf{U}_\mathbf{A}^\perp$ are equal to one. Given $\mathbf{U}_\mathbf{A}$, $\mathbf{U}_\mathbf{A}^\perp$ can be computed in deterministic $O(n(n-k)^2)$ time via the QR factorization.

We remind the reader of the Frobenius and spectral matrix norms: $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2 = \sum_{i=1}^k \sigma_i^2$ and $\|\mathbf{A}\|_2^2 = \sigma_1^2$. We will sometimes use the notation $\|\mathbf{A}\|_\xi$ to indicate that an expression holds for both $\xi = 2$ or $\xi = F$. For any two matrices \mathbf{X} and \mathbf{Y} , $\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F \leq \sqrt{\text{rank}(\mathbf{X})} \|\mathbf{X}\|_2$; $\|\mathbf{X}\mathbf{Y}\|_F \leq \|\mathbf{X}\|_F \|\mathbf{Y}\|_2$; $\|\mathbf{X}\mathbf{Y}\|_F \leq \|\mathbf{X}\|_2 \|\mathbf{Y}\|_F$. These are stronger variants of the standard submultiplicativity property $\|\mathbf{X}\mathbf{Y}\|_\xi \leq \|\mathbf{X}\|_\xi \|\mathbf{Y}\|_\xi$ and we will refer to them as spectral submultiplicativity. It follows that, if \mathbf{Q} is orthonormal, then $\|\mathbf{Q}\mathbf{X}\|_\xi \leq \|\mathbf{X}\|_\xi$ and $\|\mathbf{Y}\mathbf{Q}^\top\|_\xi \leq \|\mathbf{Y}\|_\xi$. Finally, we will make frequent use of the following two lemmas.

Lemma 17 (matrix-Pythagoras). *Let \mathbf{X} and \mathbf{Y} be two $n \times d$ matrices. If $\mathbf{X}\mathbf{Y}^\top = \mathbf{0}_{n \times n}$ or $\mathbf{X}^\top \mathbf{Y} = \mathbf{0}_{d \times d}$, then*

$$\|\mathbf{X} + \mathbf{Y}\|_\xi^2 \leq \|\mathbf{X}\|_\xi^2 + \|\mathbf{Y}\|_\xi^2.$$

Lemma 18 (Fact 6.4.12 in [4]). *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times \ell}$, and assume that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = n$. Then,*

$$(\mathbf{A}\mathbf{B})^\dagger = \mathbf{B}^\dagger \mathbf{A}^\dagger.$$

B Algorithms and Proofs of Lemmas 10 and 11

We now provide all the details of the proofs and the corresponding algorithms of Lemmas 10 and 11. Those results, which have been described in detail in [5], are slight extensions of two algorithms presented in [7], which themselves extend the original spectral sparsification result of Batson, Spielman, and Srivastava [2]. More specifically, Lemma 20 below - in some sense - generalizes Lemma 2; indeed, setting $\mathbf{V} = \mathbf{Q}^\top := \mathbf{U}$ in Lemma 20 gives Lemma 2. Lemma 19 below also describes a deterministic algorithm for sampling columns from two matrices but the goal here is to optimize different spectral properties in the sampled matrices.

In this section of the Appendix, we will slightly abuse notation by denoting with $\hat{\mathbf{S}} \in \mathbb{R}^{n \times r}$ a sampling matrix which samples columns - not rows - from matrices. We will later use $\mathbf{S} = \hat{\mathbf{S}}^\top$ to be consistent with the notation used throughout the paper.

Lemma 19 (Lemma 13 in [7]). *Let $\mathbf{V}^\top \in \mathbb{R}^{k \times n}$ and $\mathbf{B} \in \mathbb{R}^{\ell_1 \times n}$ with $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_k$. Let $r > k$. Algorithm 5 runs in $O(rk^2n + \ell_1 n)$ time and deterministically constructs a sampling matrix $\hat{\mathbf{S}} \in \mathbb{R}^{n \times r}$ and a rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$ such that,*

$$\sigma_k(\mathbf{V}^\top \hat{\mathbf{S}} \mathbf{D}) \geq 1 - \sqrt{k/r}; \quad \|\mathbf{B} \hat{\mathbf{S}} \mathbf{D}\|_F \leq \|\mathbf{B}\|_F.$$

We write $[\mathbf{D}, \hat{\mathbf{S}}] = \text{DeterministicSamplingI}(\mathbf{V}^\top, \mathbf{B}, r)$ to denote this procedure.

Input: $\mathbf{V}^T = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{k \times n}$, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n] \in \mathbb{R}^{\ell_1 \times n}$, and $r > k$.
Output: Sampling matrix $\hat{\mathbf{S}} \in \mathbb{R}^{n \times r}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.

- 1: Initialize $\mathbf{A}_0 = \mathbf{0}_{k \times k}$, $\hat{\mathbf{S}} = \mathbf{0}_{n \times r}$, and $\mathbf{D} = \mathbf{0}_{r \times r}$.
- 2: Set constants $\delta_{\mathbf{B}} = \|\mathbf{B}\|_{\text{F}}^2 (1 - \sqrt{k/r})^{-1}$; $\delta_L = 1$.
- 3: **for** $\tau = 0$ **to** $r - 1$ **do**
- 4: Let $L_\tau = \tau - \sqrt{rk}$.
- 5: Pick index $i_\tau \in \{1, 2, \dots, n\}$ and number $t_\tau > 0$ (see text for the definition of U, L):

$$U(\mathbf{b}_{i_\tau}, \delta_{\mathbf{B}}) \leq \frac{1}{t_\tau} \leq L(\mathbf{v}_{i_\tau}, \delta_L, \mathbf{A}_\tau, L_\tau).$$

- 6: Update $\mathbf{A}_{\tau+1} = \mathbf{A}_\tau + t_\tau \mathbf{v}_{i_\tau} \mathbf{v}_{i_\tau}^T$; set $\hat{\mathbf{S}}_{i_\tau, \tau+1} = 1$ and $\mathbf{D}_{\tau+1, \tau+1} = 1/\sqrt{t_\tau}$.
- 7: **end for**
- 8: Multiply all the weights in \mathbf{D} by

$$\sqrt{r^{-1}(1 - \sqrt{k/r})}.$$

- 9: **Return:** $\hat{\mathbf{S}}$ and \mathbf{D} .

Algorithm 5: DeterministicSamplingI (Lemma 19)

Algorithm 5 is a greedy technique that selects columns one at a time. To describe the algorithm in more detail, it is convenient to view the input matrices as two sets of n vectors,

$$\mathbf{V}^T = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n],$$

and

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n].$$

Given k and $r > k$, introduce the iterator $\tau = 0, 1, 2, \dots, r - 1$, and define the parameter

$$L_\tau = \tau - \sqrt{rk}.$$

For a square symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$ with eigenvalues $\lambda_1, \dots, \lambda_k$, $\mathbf{v} \in \mathbb{R}^k$ and $L \in \mathbb{R}$, define

$$\phi(L, \mathbf{A}) = \sum_{i=1}^k \frac{1}{\lambda_i - L},$$

and let $L(\mathbf{v}, \delta_L, \mathbf{A}, L)$ be defined as

$$L(\mathbf{v}, \delta_L, \mathbf{A}, L) = \frac{\mathbf{v}^T (\mathbf{A} - L' \mathbf{I}_k)^{-2} \mathbf{v}}{\phi(L', \mathbf{A}) - \phi(L, \mathbf{A})} - \mathbf{v}^T (\mathbf{A} - L' \mathbf{I}_k)^{-1} \mathbf{v},$$

where $L' = L + \delta_L = L + 1$. For a vector \mathbf{z} and scalar $\delta > 0$, define the function

$$U(\mathbf{z}, \delta) = \frac{1}{\delta} \mathbf{z}^T \mathbf{z}.$$

Input: $\mathbf{V}^\top = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{k \times n}$, $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_d] \in \mathbb{R}^{\ell_2 \times n}$, and $r > k$.
Output: Sampling matrix $\hat{\mathbf{S}} \in \mathbb{R}^{n \times r}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$.

- 1: Initialize $\mathbf{A}_0 = \mathbf{0}_{k \times k}$, $\mathbf{B}_0 = \mathbf{0}_{\ell_2 \times \ell_2}$, $\mathbf{\Omega} = \mathbf{0}_{n \times r}$, and $\hat{\mathbf{S}} = \mathbf{0}_{r \times r}$.
- 2: Set constants $\delta_{\mathbf{Q}} = (1 + \ell_2/r) \left(1 - \sqrt{k/r}\right)^{-1}$; $\delta_L = 1$.
- 3: **for** $\tau = 0$ **to** $r - 1$ **do**
- 4: Let $L_\tau = \tau - \sqrt{rk}$; $U_\tau = \delta_{\mathbf{Q}} (\tau + \sqrt{\ell_2 r})$
- 5: Pick index $i_\tau \in \{1, 2, \dots, n\}$ and number $t_\tau > 0$ (see text for the definition of U, L):

$$\hat{U}(\mathbf{q}_{i_\tau}, \delta_{\mathbf{Q}}, \mathbf{B}_\tau, U_\tau) \leq \frac{1}{t_\tau} \leq L(\mathbf{v}_{i_\tau}, \delta_L, \mathbf{A}_\tau, L_\tau).$$

- 6: Update $\mathbf{A}_{\tau+1} = \mathbf{A}_\tau + t_\tau \mathbf{v}_{i_\tau} \mathbf{v}_{i_\tau}^\top$; $\mathbf{B}_{\tau+1} = \mathbf{B}_\tau + t_\tau \mathbf{q}_{i_\tau} \mathbf{q}_{i_\tau}^\top$, and
 set $\hat{\mathbf{S}}_{i_\tau, \tau+1} = 1$, $\mathbf{D}_{\tau+1, \tau+1} = 1/\sqrt{t_\tau}$.
- 7: **end for**
- 8: Multiply all the weights in \mathbf{D} by $\sqrt{r^{-1} \left(1 - \sqrt{k/r}\right)}$.
- 9: **Return:** $\hat{\mathbf{S}}$ and \mathbf{D} .

Algorithm 6: DeterministicSamplingII (Lemma 20)

At each iteration τ , the algorithm selects i_τ , $t_\tau > 0$ for which

$$U(\mathbf{b}_{i_\tau}, \delta_{\mathbf{B}}) \leq t_\tau^{-1} \leq L(\mathbf{v}_{i_\tau}, \delta_L, \mathbf{A}_\tau, L_\tau).$$

The running time of the algorithm is dominated by the search for an index i_τ satisfying

$$U(\mathbf{b}_{i_\tau}, \delta_{\mathbf{B}}) \leq t_\tau^{-1} \leq L(\mathbf{v}_{i_\tau}, \delta^{-1}, \mathbf{A}_\tau, L_\tau)$$

(one can achieve that by exhaustive search). One needs $\phi(L, \mathbf{A})$, and hence the eigenvalues of \mathbf{A} . This takes $O(k^3)$ time, once per iteration, for a total of $O(rk^3)$. Then, for $i = 1, \dots, n$, we need to compute L for every \mathbf{v}_i . This takes $O(nk^2)$ per iteration, for a total of $O(rnk^2)$. To compute U , we need $\mathbf{b}_i^\top \mathbf{b}_i$ for $i = 1, \dots, n$, which need to be computed only once for the whole algorithm and takes $O(\ell_1 n)$. So, the total running time is $O(nrk^2 + \ell_1 n)$.

Lemma 20 (Lemma 12 in [7]). *Let $\mathbf{V}^\top \in \mathbb{R}^{k \times n}$, $\mathbf{Q} \in \mathbb{R}^{\ell_2 \times n}$, $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_k$, and $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_{\ell_2}$. Let $r > k$. Algorithm 6 runs in $O(rk^2 n + r\ell_2^2 n)$ time and deterministically constructs a sampling matrix $\hat{\mathbf{S}} \in \mathbb{R}^{n \times r}$ and a rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$ such that,*

$$\sigma_k(\mathbf{V}^\top \hat{\mathbf{S}} \mathbf{D}) \geq 1 - \sqrt{k/r}; \quad \|\mathbf{Q} \hat{\mathbf{S}} \mathbf{D}\|_2 \leq 1 + \sqrt{\ell_2/r}.$$

If $\mathbf{Q} = \mathbf{I}_n$, it runs in $O(rk^2 n)$; we write $[\mathbf{D}, \hat{\mathbf{S}}] = \text{DeterministicSamplingII}(\mathbf{V}^\top, \mathbf{Q}, r)$ for this procedure.

Algorithm 6 is similar to Algorithm 5; we only need to define the function \hat{U} . For a square symmetric matrix $\mathbf{B} \in \mathbb{R}^{\ell_2 \times \ell_2}$ with eigenvalues $\lambda_1, \dots, \lambda_{\ell_2}$, $\mathbf{q} \in \mathbb{R}^{\ell_2}$, $u \in \mathbb{R}$, define: $\hat{\phi}(u, \mathbf{B}) = \sum_{i=1}^{\ell_2} \frac{1}{u - \lambda_i}$, and let $\hat{U}(\mathbf{q}, \delta_{\mathbf{Q}}, \mathbf{B}, u)$ be defined as $\hat{U}(\mathbf{q}, \delta_{\mathbf{Q}}, \mathbf{B}, u) = \frac{\mathbf{q}^\top (\mathbf{B} - u' \mathbf{I}_{\ell_2})^{-2} \mathbf{q}}{\hat{\phi}(u, \mathbf{B}) - \hat{\phi}(u', \mathbf{B})} - \mathbf{q}^\top (\mathbf{B} - u' \mathbf{I}_{\ell_2})^{-1} \mathbf{q}$, where $u' = u + \delta_{\mathbf{Q}} = u + (1 + \ell_2/r) \left(1 - \sqrt{k/r}\right)^{-1}$. The running time of the algorithm is $O(nrk^2 + nr\ell_2^2)$.

B.1 Proof of Lemma 10

We first restate the lemma.

Lemma 21 (Restatement of Lemma 10). *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\Psi \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_{\mathbf{Y}}$, and ρ_{Ψ} . Given $r > \rho_{\mathbf{Y}}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + T_{\text{SVD}}(\Psi) + O(rn(\rho_{\mathbf{Y}}^2 + \rho_{\Psi}^2))$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{DSY}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{DSY})^\dagger\|_2 < \frac{1}{1 - \sqrt{\rho_{\mathbf{Y}}/r}} \|\mathbf{Y}^\dagger\|_2; \quad \|\mathbf{DS\Psi}\|_2 < \left(1 + \sqrt{\frac{\rho_{\Psi}}{r}}\right) \|\Psi\|_2.$$

If $\Psi = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_{\mathbf{Y}}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{Y}, \Psi, r)$ to denote such a deterministic procedure.

Proof. Let the SVD of $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ is $\mathbf{Y} = \mathbf{U}_{\mathbf{Y}} \Sigma_{\mathbf{Y}} \mathbf{V}_{\mathbf{Y}}^T$, with $\mathbf{U}_{\mathbf{Y}} \in \mathbb{R}^{n \times \rho_{\mathbf{Y}}}$, $\Sigma_{\mathbf{Y}} \in \mathbb{R}^{\rho_{\mathbf{Y}} \times \rho_{\mathbf{Y}}}$, $\mathbf{V}_{\mathbf{Y}} \in \mathbb{R}^{\ell_1 \times \rho_{\mathbf{Y}}}$. Let the SVD of $\Psi \in \mathbb{R}^{n \times \ell_2}$ is $\Psi = \mathbf{U}_{\Psi} \Sigma_{\Psi} \mathbf{V}_{\Psi}^T$, with $\mathbf{U}_{\Psi} \in \mathbb{R}^{n \times \rho_{\Psi}}$, $\Sigma_{\Psi} \in \mathbb{R}^{\rho_{\Psi} \times \rho_{\Psi}}$, and $\mathbf{V}_{\Psi} \in \mathbb{R}^{\ell_2 \times \rho_{\Psi}}$. Let

$$[\mathbf{D}, \hat{\mathbf{S}}] = \text{DeterministicSamplingII}(\mathbf{U}_{\mathbf{Y}}^T, \mathbf{U}_{\Psi}^T, r).$$

By Lemma 20,

$$\sigma_{\min}(\mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D}) \geq \left(1 - \sqrt{\rho_{\mathbf{Y}}/r}\right),$$

which implies

$$\|(\mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \leq \left(1 - \sqrt{\rho_{\mathbf{Y}}/r}\right)^{-1},$$

and

$$\text{rank}(\mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D}) = \rho_{\mathbf{Y}}.$$

Also,

$$\|\mathbf{U}_{\Psi}^T \hat{\mathbf{S}} \mathbf{D}\|_2 \leq \left(1 + \sqrt{\rho_{\Psi}/r}\right)$$

because

$$\sigma_{\max}(\mathbf{U}_{\Psi}^T \hat{\mathbf{S}} \mathbf{D}) \leq \left(1 + \sqrt{\rho_{\Psi}/r}\right).$$

Thus,

$$\begin{aligned} \|(\mathbf{Y}^T \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 &= \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}} \mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \\ &\stackrel{(a)}{=} \|(\mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D})^\dagger (\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 \\ &\leq \|(\mathbf{U}_{\mathbf{Y}}^T \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 \\ &\leq \left(1 - \sqrt{\rho_{\mathbf{Y}}/r}\right)^{-1} \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 \\ &= \left(1 - \sqrt{\rho_{\mathbf{Y}}/r}\right)^{-1} \|(\mathbf{Y}^T)^\dagger\|_2 \end{aligned}$$

(a) uses Lemma 18. To obtain the first inequality in the lemma we need to take $\mathbf{S} = \hat{\mathbf{S}}^T$ and observe that $\|(\mathbf{Y}^T \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 = \|(\mathbf{DSY})^\dagger\|_2$, and $\|(\mathbf{Y}^T)^\dagger\|_2 = \|\mathbf{Y}^\dagger\|_2$. We now prove the second inequality in the lemma,

$$\|\Psi^T \hat{\mathbf{S}} \mathbf{D}\|_2 = \|\mathbf{V}_{\Psi} \Sigma_{\Psi} \mathbf{U}_{\Psi}^T \hat{\mathbf{S}} \mathbf{D}\|_2 \leq \|\mathbf{V}_{\Psi} \Sigma_{\Psi}\|_2 \|\mathbf{U}_{\Psi}^T \hat{\mathbf{S}} \mathbf{D}\|_2 = \|\Psi^T\|_2 \|\mathbf{U}_{\Psi}^T \hat{\mathbf{S}} \mathbf{D}\|_2 \leq \|\Psi^T\|_2 \left(1 + \sqrt{\rho_{\Psi}/r}\right).$$

To obtain the second inequality in the lemma we need to take $\mathbf{S} = \hat{\mathbf{S}}^T$ and use $\|\Psi^T \hat{\mathbf{S}} \mathbf{D}\|_2 = \|\mathbf{DS\Psi}\|_2$, and $\|\Psi^T\|_2 = \|\Psi\|_2$. \blacksquare

B.2 Proof of Lemma 11

We first restate the lemma.

Lemma 22 (Restatement of Lemma 11). *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\Psi \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_{\mathbf{Y}}$, and ρ_{Ψ} . Given $r > \rho_{\mathbf{Y}}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_{\mathbf{Y}}^2 + \ell_2 n)$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{DSY}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{DSY})^\dagger\|_2 < \frac{1}{1 - \sqrt{\rho_{\mathbf{Y}}/r}} \|\mathbf{Y}^\dagger\|_2; \quad \|\mathbf{DS}\Psi\|_{\text{F}} \leq \|\Psi\|_{\text{F}}.$$

If $\Psi = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_{\mathbf{Y}}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleFrobeniusSampling}(\mathbf{Y}, \Psi, r)$ to denote such a deterministic procedure.

Proof. Let the SVD of $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ is $\mathbf{Y} = \mathbf{U}_{\mathbf{Y}} \Sigma_{\mathbf{Y}} \mathbf{V}_{\mathbf{Y}}^{\text{T}}$, with $\mathbf{U}_{\mathbf{Y}} \in \mathbb{R}^{n \times \rho_{\mathbf{Y}}}$, $\Sigma_{\mathbf{Y}} \in \mathbb{R}^{\rho_{\mathbf{Y}} \times \rho_{\mathbf{Y}}}$, $\mathbf{V}_{\mathbf{Y}} \in \mathbb{R}^{\ell_1 \times \rho_{\mathbf{Y}}}$. Let the SVD of $\Psi \in \mathbb{R}^{n \times \ell_2}$ is $\Psi = \mathbf{U}_{\Psi} \Sigma_{\Psi} \mathbf{V}_{\Psi}^{\text{T}}$, with $\mathbf{U}_{\Psi} \in \mathbb{R}^{n \times \rho_{\Psi}}$, $\Sigma_{\Psi} \in \mathbb{R}^{\rho_{\Psi} \times \rho_{\Psi}}$, and $\mathbf{V}_{\Psi} \in \mathbb{R}^{\ell_2 \times \rho_{\Psi}}$. Let

$$[\mathbf{D}, \hat{\mathbf{S}}] = \text{DeterministicSamplingI}(\mathbf{U}_{\mathbf{Y}}^{\text{T}}, \Psi^{\text{T}}, r).$$

By Lemma 19,

$$\sigma_{\min}(\mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D}) \geq \left(1 - \sqrt{\rho_{\mathbf{Y}}/r}\right),$$

which implies

$$\|(\mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \leq (1 - \sqrt{\rho_{\mathbf{Y}}/r})^{-1},$$

and

$$\text{rank}(\mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D}) = \rho_{\mathbf{Y}}.$$

Also,

$$\|\Psi^{\text{T}} \hat{\mathbf{S}} \mathbf{D}\|_{\text{F}} \leq \|\Psi^{\text{T}}\|_{\text{F}},$$

which by taking $\mathbf{S} = \hat{\mathbf{S}}^{\text{T}}$ gives the second inequality in the lemma,

$$\|\mathbf{DS}\Psi\|_{\text{F}} \leq \|\Psi\|_{\text{F}}.$$

Now we prove the first inequality in the lemma,

$$\begin{aligned} \|(\mathbf{Y}^{\text{T}} \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 &= \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}} \mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \stackrel{(a)}{=} \|(\mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D})^\dagger (\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 \leq \|(\mathbf{U}_{\mathbf{Y}}^{\text{T}} \hat{\mathbf{S}} \mathbf{D})^\dagger\|_2 \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 \\ &\leq (1 - \sqrt{\rho_{\mathbf{Y}}/r})^{-1} \|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2. \end{aligned}$$

(a) uses Lemma 18. To conclude, use $\|(\mathbf{Y}^{\text{T}} \mathbf{S})^\dagger\|_2 = \|(\mathbf{DSY})^\dagger\|_2$; $\|(\mathbf{V}_{\mathbf{Y}} \Sigma_{\mathbf{Y}})^\dagger\|_2 = \|(\mathbf{Y}^{\text{T}})^\dagger\|_2 = \|(\mathbf{Y})^\dagger\|_2$. ■