arXiv:1102.3975v2 [stat.ML] 25 Feb 2011

Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection

Abhimanyu Das
University of Southern California
abhimand@usc.edu

David Kempe*
University of Southern California
dkempe@usc.edu

Abstract

We study the problem of selecting a subset of k random variables from a large set, in order to obtain the best linear prediction of another variable of interest. This problem can be viewed in the context of both feature selection and sparse approximation. We analyze the performance of widely used greedy heuristics, using insights from the maximization of submodular functions and spectral analysis. We introduce the *submodularity ratio* as a key quantity to help understand why greedy algorithms perform well even when the variables are highly correlated. Using our techniques, we obtain the strongest known approximation guarantees for this problem, both in terms of the submodularity ratio and the smallest k-sparse eigenvalue of the covariance matrix.

We further demonstrate the wide applicability of our techniques by analyzing greedy algorithms for the dictionary selection problem, and significantly improve the previously known guarantees. Our theoretical analysis is complemented by experiments on real-world and synthetic data sets; the experiments show that the submodularity ratio is a stronger predictor of the performance of greedy algorithms than other spectral parameters.

1 Introduction

We analyze algorithms for the following important *Subset Selection* problem: select a subset of k variables from a given set of n observation variables which, taken together, "best" predict another variable of interest. This problem has many applications ranging from feature selection, sparse learning and dictionary selection in machine learning, to sparse approximation and compressed sensing in signal processing. From a machine learning perspective, the variables could be features or observable attributes of a phenomenon, and we wish to predict the phenomenon using only a small subset from the high-dimensional feature space. In signal processing, the variables could correspond to a collection of dictionary vectors, and the goal is to parsimoniously represent another (output) vector. For many practitioners, the prediction model of choice is linear regression, and the goal is to obtain a linear model using a small subset of variables, to minimize the mean square prediction error or, equivalently, maximize the squared multiple correlation R^2 [6].

Thus, we formulate the Subset Selection problem for regression as follows: Given the (normalized) covariances between n variables X_i (which can in principle be observed) and a variable Z (which is to be predicted), select a subset of $k \ll n$ of the variables X_i and a linear prediction function of Z from the selected X_i that maximizes the R^2 fit. (A formal definition is given in Section 2.) The covariances are usually obtained empirically from detailed past observations of the variable values.

*Supported in part by NSF CAREER award 0545855, and NSF grant dddas-tmrp 0540420

The above formulation is known [2] to be equivalent to the problem of *sparse approximation* over dictionary vectors: the input consists of a dictionary of n feature vectors $\mathbf{x_i} \in \mathbb{R}^m$, along with a target vector $\mathbf{z} \in \mathbb{R}^m$, and the goal is to select at most k vectors whose linear combination best approximates $\mathbf{z}$. The pairwise covariances of the previous formulation are then exactly the inner products of the dictionary vectors.[1]

Our problem formulation appears somewhat similar to the problem of sparse recovery [17, 18, 19, 1]; however, note that in sparse recovery, it is generally assumed that the prediction vector is truly (almost) k-sparse, and the aim is to recover the exact coefficients of this truly sparse solution. However, finding a sparse solution is a well-motivated problem even if the true solution is not sparse. Even then, running subset selection to find a sparse approximation to the correct solution helps to reduce cost and model complexity.

This problem is $\mathbf{NP}$-hard [11], so no polynomial-time algorithms are known to solve it optimally for all inputs. Two approaches are frequently used for approximating such problems: greedy algorithms [10, 14, 5, 17] and convex relaxation schemes [13, 1, 15, 4]. For our formulation, a disadvantage of convex relaxation techniques is that they do not provide explicit control over the target sparsity level k of the solution; additional effort is needed to tune the regularization parameter.

A simpler and more intuitive approach, widely used in practice for subset selection problems (for example, it is implemented in all commercial statistics packages), is to use greedy algorithms, which iteratively add or remove variables based on simple measures of fit with Z. Two of the most well-known and widely used greedy algorithms are the subject of our analysis: Forward Regression [10] and Orthogonal Matching Pursuit [14]. (These algorithms are defined formally in Section 2).

So far, the theoretical bounds on such greedy algorithms have been unable to explain why they perform well in practice for most subset selection problem instances. Most previous results for greedy subset selection algorithms [5, 14, 2] have been based on coherence of the input data, i.e., the maximum correlation μ between any pair of variables. Small coherence is an extremely strong condition, and the bounds usually break down when the coherence is $\omega(1/k)$. On the other hand, most bounds for greedy and convex relaxation algorithms for sparse recovery are based on a weaker sparse-eigenvalue or Restricted Isometry Property (RIP) condition [18, 17, 9, 20, 1]. However, these results apply to a different objective: minimizing the difference between the actual and estimated coefficients of a sparse vector. Simply extending these results to the subset selection problem adds a dependence on the largest k-sparse eigenvalue and only leads to weak additive bounds. More importantly, all the above results rely on spectral conditions that suffer from an inability to explain the performance of the algorithms for near-singular matrices.

Eigenvalue-based bounds fail to explain an observation of many experiments (including ours in Section 5): greedy algorithms often perform very well, even for near-singular input matrices. Our results begin to explain these observations by proving that the performance of many algorithms does not really depend on how singular the covariance matrix is, but rather on how far the R^2 measure deviates from submodularity on the given input. We formalize this intuition by defining a measure of "approximate submodularity" which we term *submodularity ratio*. We prove that whenever the submodularity ratio is bounded away from 0, the R^2 objective is "reasonably close" to submodular, and Forward Regression gives a constant-factor approximation. This significantly generalizes a recent result of Das and Kempe [2], who had identified a strong condition termed "absence of conditional suppressors" which ensures that the R^2 objective is actually submodular.

An analysis based on the submodularity ratio does relate with traditional spectral bounds, in that the ratio is always lower-bounded by the smallest k-sparse eigenvalue of C (though it can be significantly larger

[1] For this reason, the dimension m of the feature vectors only affects the problem indirectly, via the accuracy of the estimated covariance matrix.

when the predictor variable is not badly aligned with the eigenspace of small eigenvalues). In particular, we also obtain multiplicative approximation guarantees for both Forward Regression and Orthogonal Matching Pursuit, whenever the smallest k-sparse eigenvalue of C is bounded away from 0, significantly strengthening past known bounds on their performance.

An added benefit of our framework is that we obtain much tighter theoretical performance guarantees for greedy algorithms for dictionary selection [8]. In the *dictionary selection problem*, we are given s target vectors, and a candidate set V of feature vectors. The goal is to select a set $D \subset V$ of at most d feature vectors, which will serve as a *dictionary* in the following sense. For each of several target vectors, the best $k < d$ vectors from D will be selected and used to achieve a good R^2 fit; the goal is to maximize the average R^2 fit for all of these vectors. (A formal definition is given in Section 2.) This problem of finding a dictionary of basis functions for sparse representation of signals has several applications in machine learning and signal processing. Krause and Cevher [8] showed that greedy algorithms for dictionary selection can perform well in many instances, and proved additive approximation bounds for two specific algorithms, $\mathrm{SDS_{MA}}$ and $\mathrm{SDS_{OMP}}$ (defined in Section 4). Our approximate submodularity framework allows us to obtain much stronger multiplicative guarantees without much extra effort.

Our theoretical analysis is complemented by experiments comparing the performance of the greedy algorithms and a baseline convex-relaxation algorithm for subset selection on two real-world data sets and a synthetic data set. More importantly, we evaluate the submodularity ratio of these data sets and compare it with other spectral parameters: while the input covariance matrices are close to singular, the submodularity ratio actually turns out to be significantly larger. Thus, our theoretical results can begin to explain why, in many instances, greedy algorithms perform well in spite of the fact that the data may have high correlations.

Our main contributions can be summarized as follows:

1. We introduce the notion of the submodularity ratio as a much more accurate predictor of the performance of greedy algorithms than previously used parameters.

2. We obtain the strongest known theoretical performance guarantees for greedy algorithms for subset selection. In particular, we show (in Section 3) that the Forward Regression and OMP algorithms are within a $1 - e^{-\gamma}$ factor and $1 - e^{-(\gamma \cdot \lambda_{\min})}$ factor of the optimal solution, respectively (where the γ and λ terms are appropriate submodularity and sparse-eigenvalue parameters).

3. We obtain the strongest known theoretical guarantees for algorithms for dictionary selection, improving on the results of [8]. In particular, we show (in Section 4) that the $\mathrm{SDS_{MA}}$ algorithm is within a factor $\frac{\gamma}{\lambda_{\max}}(1 - \frac{1}{e})$ of optimal.

4. We evaluate our theoretical bounds for subset selection by running greedy and L1-relaxation algorithms on real-world and synthetic data, and show how the various submodular and spectral parameters correlate with the performance of the algorithms in practice.

1.1 Related Work

As mentioned previously, there has been a lot of recent interest in greedy and convex relaxation techniques for the sparse recovery problems, both in the noiseless and noisy setting. For L1 relaxation techniques, Tropp [15] showed conditions based on the coherence (i.e., the maximum correlation between any pair of variables) of the dictionary that guaranteed near-optimal recovery of a sparse signal. In [1, 4], it was shown that if the target signal is truly sparse, and the dictionary obeys a restricted isometry property (RIP), then L1 relaxation can almost exactly recover the true sparse signal. Other results [19, 20] also prove conditions under which

L1 relaxation can recover a sparse signal. Though related, the above results are not directly applicable to our subset selection formulation, since the goal in sparse recovery is to recover the true coefficients of the sparse signal, as opposed to our problem of minimizing the prediction error of an arbitrary signal subject to a specified sparsity level.

For greedy sparse recovery, Zhang [17, 18] and Lozano et al. [9] provided conditions based on sparse eigenvalues under which Forward Regression and Forward-Backward Regression can recover a sparse signal. As with the L1 results for sparse recovery, the objective function analyzed in these papers is somewhat different from that in our subset selection formulation; furthermore, these results are intended mainly for the case when the predictor variable is truly sparse. Simply extending these results to our problem formulation gives weaker, additive bounds and requires stronger conditions than our results.

The papers by Das and Kempe [2], Gilbert et al. [5] and Tropp et al. [16, 14] analyzed greedy algorithms using the same subset selection formulation presented in this work. In particular, they obtained a $1+\Theta(\mu^2 k)$ multiplicative approximation guarantee for the mean square error objective and a $1-\Theta(\mu k)$ guarantee for the R^2 objective, whenever the coherence μ of the dictionary is $O(1/k)$. These results are thus weaker than those presented here, since they do not apply to instances with even moderate correlations of $\omega(1/k)$.

Other analysis of greedy methods includes the work of Natarajan [11], which proved a bicriteria approximation bound for minimizing the number of vectors needed to achieve a given prediction error.

As mentioned earlier, the paper by Krause and Cevher [8] analyzed greedy algorithms for the dictionary selection problem, which generalizes subset selection to prediction of multiple variables. They too use a notion of approximate submodularity to provide additive approximation guarantees. Since their analysis is for a more general problem than subset selection, applying their results directly to the subset selection problem predictably gives much weaker bounds than those presented in this paper for subset selection. Furthermore, even for the general dictionary selection problem, our techniques can be used to significantly improve their analysis of greedy algorithms and obtain tighter multiplicative approximation bounds (as shown in Section 4).

In general, we note that the performance bounds for greedy algorithms derived using the coherence parameter are usually the weakest, followed by those using the Restricted Isometry Property, then those using sparse eigenvalues, and finally those using the submodularity ratio. (We show an empirical comparison of these parameters in Section 5.)

2 Preliminaries

The goal in subset selection is to estimate a *predictor variable* Z using linear regression on a small subset from the set of *observation variables* $V = \{X_1, \ldots, X_n\}$. We use $\mathrm{Var}(X_i)$, $\mathrm{Cov}(X_i, X_j)$ and $\rho(X_i, X_j)$ to denote the variance, covariance and correlation of random variables, respectively. By appropriate normalization, we can assume that all the random variables have expectation 0 and variance 1. The matrix of covariances between the X_i and X_j is denoted by C, with entries $c_{i,j} = \mathrm{Cov}(X_i, X_j)$. Similarly, we use $\mathbf{b}$ to denote the covariances between Z and the X_i, with entries $b_i = \mathrm{Cov}(Z, X_i)$. Formally, the *Subset Selection* problem can now be stated as follows:

Definition 2.1 (Subset Selection) *Given pairwise covariances among all variables, as well as a parameter k, find a set $S \subset V$ of at most k variables X_i and a linear predictor $Z' = \sum_{i \in S} \alpha_i X_i$ of Z, maximizing the* squared multiple correlation *[3, 6]*

$$R^2_{Z,S} = \frac{\mathrm{Var}(Z) - \mathrm{E}[(Z - Z')^2]}{\mathrm{Var}(Z)}.$$

R^2 is a widely used measure for the goodness of a statistical fit; it captures the fraction of the variance of Z explained by variables in S. Because we assumed Z to be normalized to have variance 1, it simplifies to $R^2_{Z,S} = 1 - \mathrm{E}\left[(Z - Z')^2\right]$.

For a set S, we use C_S to denote the submatrix of C with row and column set S, and $\mathbf{b}_S$ to denote the vector with only entries b_i for $i \in S$. For notational convenience, we frequently do not distinguish between the index set S and the variables $\{X_i \mid i \in S\}$. Given the subset S of variables used for prediction, the optimal regression coefficients α_i are well known to be $\mathbf{a}_S = (\alpha_i)_{i \in S} = C_S^{-1} \cdot \mathbf{b}_S$ (see, e.g., [6]), and hence $R^2_{Z,S} = \mathbf{b}_S^T C_S^{-1} \mathbf{b}_S$. Thus, the subset selection problem can be phrased as follows: Given C, $\mathbf{b}$, and k, select a set S of at most k variables to maximize $R^2_{Z,S} = \mathbf{b}_S^T (C_S^{-1}) \mathbf{b}_S$.[2]

The dictionary selection problem generalizes the subset selection problem by considering s predictor variables $Z_1, Z_2, \ldots, Z_s$. The goal is to select a dictionary D of d observation variables, to optimize the average R^2 fit for the Z_i using at most k vectors from D for each. Formally, the Dictionary Selection problem is defined as follows:

Definition 2.2 (Dictionary Selection) *Given all pairwise covariances among the Z_j and X_i variables, as well as parameters d and k, find a set D of at most d variables from $\{X_1, \ldots, X_n\}$ maximizing*

$$F(D) = \sum_{j=1}^{s} \max_{S \subset D, |S| = k} R^2_{Z_j,S}.$$

Many of our results are phrased in terms of eigenvalues of the covariance matrix C and its submatrices. Since covariance matrices are positive semidefinite, their eigenvalues are real and non-negative [6]. For any positive semidefinite $n \times n$ matrix A, we denote its eigenvalues by $\lambda_{\min}(A) = \lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A) = \lambda_{\max}(A)$. We use $\lambda_{\min}(C, k) = \min_{S:|S|=k} \lambda_{\min}(C_S)$ to refer to the smallest eigenvalue of any $k \times k$ submatrix of C (i.e., the smallest k-sparse eigenvalue), and similarly $\lambda_{\max}(C, k) = \max_{S:|S|=k} \lambda_{\max}(C_S)$.[3] We also use $\kappa(C, k)$ to denote the largest condition number (the ratio of the largest and smallest eigenvalue) of any $k \times k$ submatrix of C. This quantity is strongly related to the Restricted Isometry Property in [1]. We also use $\mu(C) = \max_{i \neq j} |c_{i,j}|$ to denote the *coherence*, i.e., the maximum absolute pairwise correlation between the X_i variables. Recall the L_2 vector and matrix norms: $\|\mathbf{x}\|_2 = \sqrt{\sum_i |x_i|^2}$, and $\|A\|_2 = \lambda_{\max}(A) = \max_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2$. We also use $\|\mathbf{x}\|_0 = |\{i \mid x_i \neq 0\}|$ to denote the sparsity of a vector $\mathbf{x}$.

The part of a variable Z that is not correlated with the X_i for all $i \in S$, i.e., the part that cannot be explained by the X_i, is called the *residual* (see [3]), and defined as $\mathrm{Res}(Z, S) = Z - \sum_{i \in S} \alpha_i X_i$.

2.1 Submodularity Ratio

We introduce the notion of submodularity ratio for a general set function, which captures "how close" to submodular the function is. We first define it for arbitrary set functions, and then show the specialization for the R^2 objective.

Definition 2.3 (Submodularity Ratio) *Let f be a non-negative set function. The* submodularity ratio *of f with respect to a set U and a parameter $k \geq 1$ is*

$$\gamma_{U,k}(f) = \min_{L \subseteq U, S: |S| \leq k, S \cap L = \emptyset} \frac{\sum_{x \in S} f(L \cup \{x\}) - f(L)}{f(L \cup S) - f(L)}.$$

[2] We assume throughout that C_S is non-singular. For some of our results, an extension to singular matrices is possible using the Moore-Penrose generalized inverse.

[3] Computing $\lambda_{\min}(C, k)$ is NP-hard. In Appendix A we describe how to efficiently approximate the values for some scenarios.

Thus, it captures how much more f can increase by adding any subset S of size k to L, compared to the combined benefits of adding its individual elements to L.

If f is specifically the R^2 objective defined on the variables X_i, then we omit f and simply define

$$\gamma_{U,k} = \min_{L \subseteq U, S:|S| \leq k, S \cap L = \emptyset} \frac{\sum_{i \in S}(R^2_{Z,L \cup \{X_i\}} - R^2_{Z,L})}{R^2_{Z,S \cup L} - R^2_{Z,L}} = \min_{L \subseteq U, S:|S| \leq k, S \cap L = \emptyset} \frac{(\mathbf{b}^L_S)^T \mathbf{b}^L_S}{(\mathbf{b}^L_S)^T (C^L_S)^{-1} \mathbf{b}^L_S},$$

where C^L and $\mathbf{b}^L$ are the normalized covariance matrix and normalized covariance vector corresponding to the set $\{\mathrm{Res}(X_1, L), \mathrm{Res}(X_2, L), \ldots, \mathrm{Res}(X_n, L)\}$.

It can be easily shown that f is submodular if and only if $\gamma_{U,k} \geq 1$, for all U and k. For the purpose of subset selection, it is significant that the submodularity ratio can be bounded in terms of the smallest sparse eigenvalue, as shown in the following lemma:

Lemma 2.4 $\gamma_{U,k} \geq \lambda_{\min}(C, k + |U|) \geq \lambda_{\min}(C)$.

For all our analysis in this paper, we will use $|U| = k$, and hence $\gamma_{U,k} \geq \lambda_{\min}(C, 2k)$. Thus, the smallest $2k$-sparse eigenvalue is a lower bound on this submodularity ratio; as we show later, it is often a weak lower bound.

Before proving Lemma 2.4, we first introduce two lemmas that relate the eigenvalues of normalized covariance matrices with those of its submatrices.

Lemma 2.5 *Let C be the covariance matrix of n zero-mean random variables $X_1, X_2, \ldots, X_n$, each of which has variance at most 1. Let C_ρ be the corresponding correlation matrix of the n random variables, that is, C_ρ is the covariance matrix of the variables after they are normalized to have unit variance. Then $\lambda_{\min}(C) \leq \lambda_{\min}(C_\rho)$.*

Proof. Since C_ρ is obtained by normalizing the variables such that they have unit variance, we get $C_\rho = D^T C D$, where D is a diagonal matrix with diagonal entries $d_i = \frac{1}{\sqrt{\mathrm{Var}(X_i)}}$.

Since both C_ρ and C are positive semidefinite, we can perform Cholesky factorization to get lower-triangular matrices A_ρ and A such that $C = AA^T$ and $C_\rho = A_\rho A_\rho^T$. Hence $A_\rho = D^T A$.

Let $\sigma_{\min}(A)$ and $\sigma_{\min}(A_\rho)$ denote the smallest singular values of A and A_ρ, respectively. Also, let $\mathbf{v}$ be the singular vector corresponding to $\sigma_{\min}(A_\rho)$. Then,

$$\|A\mathbf{v}\|_2 = \|D^{-1} A_\rho \mathbf{v}\|_2 \leq \|D^{-1}\|_2 \|A_\rho \mathbf{v}\|_2 = \sigma_{\min}(A_\rho) \|D^{-1}\|_2 \leq \sigma_{\min}(A_\rho),$$

where the last inequality follows since

$$\|D^{-1}\|_2 = \max_i \frac{1}{d_i} = \max_i \sqrt{\mathrm{Var}(X_i)} \leq 1.$$

Hence, by the Courant-Fischer theorem, $\sigma_{\min}(A) \leq \sigma_{\min}(A_\rho)$, and consequently, $\lambda_{\min}(C) \leq \lambda_{\min}(C_\rho)$. ■

Lemma 2.6 *Let $\lambda_{\min}(C)$ be the smallest eigenvalue of the covariance matrix C of n random variables $X_1, X_2, \ldots, X_n$, and $\lambda_{\min}(C')$ be the smallest eigenvalue of the $(n-1) \times (n-1)$ covariance matrix C' corresponding to the $n-1$ random variables $\mathrm{Res}(X_1, X_n), \ldots, \mathrm{Res}(X_{n-1}, X_n)$. Then $\lambda_{\min}(C) \leq \lambda_{\min}(C')$.*

Proof. Let λ_i and λ'_i denote the eigenvalues of C and C' respectively. Also, let $c'_{i,j}$ denote the entries of C'. Using the definition of the residual, we get that

$$\begin{aligned} c'_{i,j} &= \mathrm{Cov}(\mathrm{Res}(X_i, X_n), \mathrm{Res}(X_j, X_n)) &= c_{i,j} - \tfrac{c_{i,n}c_{j,n}}{c_{n,n}}, \\ c'_{i,i} &= \mathrm{Var}(\mathrm{Res}(X_i, X_n)) &= c_{i,i} - \tfrac{c_{i,n}^2}{c_{n,n}}. \end{aligned}$$

Defining $D = \frac{1}{c_{n,n}} \cdot [c_{1,n}, c_{2,n}, \ldots, c_{n-1,n}]^T \cdot [c_{1,n}, c_{2,n}, \ldots, c_{n-1,n}]$, we can write $C_{\{1,\ldots,n-1\}} = C' + D$. To prove $\lambda_1 \leq \lambda'_1$, let $\mathbf{e}' = [e'_1, \ldots, e'_{n-1}]^T$ be the eigenvector of C' corresponding to the eigenvalue λ'_1, and consider the vector $\mathbf{e} = [e'_1, e'_2, \ldots, e'_{n-1}, -\frac{1}{c_{n,n}} \sum_{i=1}^{n-1} e'_i c_{i,n}]^T$. Then, $C \cdot \mathbf{e} = [\begin{smallmatrix}\mathbf{y}\\0\end{smallmatrix}]$, where

$$\begin{aligned} \mathbf{y} &= -\frac{1}{c_{n,n}} \sum_{i=1}^{n-1} e'_i c_{i,n} [c_{1,n}, c_{2,n}, \ldots, c_{n-1,n}]^T + C_{\{1,\ldots,n-1\}} \cdot \mathbf{e}' \\ &= -\frac{1}{c_{n,n}} \sum_{i=1}^{n-1} e'_i c_{i,n} [c_{1,n}, c_{2,n}, \ldots, c_{n-1,n}]^T + D \cdot \mathbf{e}' + C' \cdot \mathbf{e}' \\ &= C' \cdot \mathbf{e}'. \end{aligned}$$

Thus, $C \cdot \mathbf{e} = [\lambda'_1 e'_1, \lambda'_1 e'_2, \ldots, \lambda'_1 e'_{n-1}, 0]^T = \lambda'_1 [e'_1, e'_2, \ldots, e'_{n-1}, 0]^T \leq \lambda'_1 \|\mathbf{e}\|_2$, which by Rayleigh-Ritz bounds implies that $\lambda_1 \leq \lambda'_1$. ∎

Using the above two lemmas, we now prove Lemma 2.4.

Proof of Lemma 2.4. Since

$$\frac{(\mathbf{b}_S^L)^T (C_S^L)^{-1} \mathbf{b}_S^L}{(\mathbf{b}_S^L)^T \mathbf{b}_S^L} \;\leq\; \max_{\mathbf{x}} \frac{\mathbf{x}^T (C_S^L)^{-1} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \;=\; \lambda_{\max}((C_S^L)^{-1}) \;=\; \frac{1}{\lambda_{\min}(C_S^L)},$$

we can use Definition 2.3 to obtain that

$$\gamma_{U,k} \geq \min_{(L \subseteq U, S : |S| \leq k, S \cap L = \emptyset)} \lambda_{\min}(C_S^L).$$

Next, we relate $\lambda_{\min}(C_S^L)$ with $\lambda_{\min}(C_{L \cup S})$, using repeated applications of Lemmas 2.5 and 2.6. Let $L = \{X_1, \ldots, X_\ell\}$; for each i, define $L_i = \{X_1, \ldots, X_i\}$, and let $C^{(i)}$ be the covariance matrix of the random variables $\{\mathrm{Res}(X, L \setminus L_i) \mid X \in S \cup L_i\}$, and $C_\rho^{(i)}$ the covariance matrix after normalizing all its variables to unit variance. Then, Lemma 2.5 implies that for each i, $\lambda_{\min}(C^{(i)}) \leq \lambda_{\min}(C_\rho^{(i)})$, and Lemma 2.6 shows that $\lambda_{\min}(C_\rho^{(i)}) \leq \lambda_{\min}(C^{(i-1)})$ for each $i > 0$. Combining these inequalities inductively for all i, we obtain that

$$\lambda_{\min}(C_S^L) \;=\; \lambda_{\min}(C_\rho^{(0)}) \;\geq\; \lambda_{\min}(C^{(\ell)}) \;=\; \lambda_{\min}(C_{L \cup S}) \;\geq\; \lambda_{\min}(C, |L \cup S|).$$

Finally, since $|S| \leq k$ and $L \subseteq U$, we obtain $\gamma_{U,k} \geq \lambda_{\min}(C, k + |U|)$. ∎

3 Algorithms Analysis

We now present theoretical performance bounds for Forward Regression and Orthogonal Matching Pursuit, which are widely used in practice. We also analyze the Oblivious algorithm: one of the simplest greedy algorithms for subset selection. Throughout this section, we use $\mathrm{OPT} = \max_{S:|S|=k} R^2_{Z,S}$ to denote the optimum R^2 value achievable by any set of size k.

3.1 Forward Regression

We first provide approximation bounds for Forward Regression, which is the standard algorithm used by many researchers in medical, social, and economic domains.[4]

Definition 3.1 (Forward Regression) *The* Forward Regression *(also called* Forward Selection*) algorithm for subset selection selects a set S of size k iteratively as follows:*

1: *Initialize $S_0 = \emptyset$.*
2: ***for** each iteration $i + 1$ **do***
3: *Let X_m be a variable maximizing $R^2_{Z,S_i\cup\{X_m\}}$, and set $S_{i+1} = S_i \cup \{X_m\}$.*
4: *Output S_k.*

Our main result is the following theorem.

Theorem 3.2 *The set S^{FR} selected by forward regression has the following approximation guarantees:*

$$\begin{aligned} R^2_{Z,S^{FR}} &\geq (1 - e^{-\gamma_{S^{FR},k}}) \cdot OPT \\ &\geq (1 - e^{-\lambda_{\min}(C,2k)}) \cdot OPT \\ &\geq (1 - e^{-\lambda_{\min}(C,k)}) \cdot \Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)}) \cdot OPT. \end{aligned}$$

Before proving the theorem, we first begin with a general lemma that bounds the amount by which the R^2 value of a set and the sum of R^2 values of its elements can differ.

Lemma 3.3 $\frac{1}{\lambda_{\max}(C)} \sum_{i=1}^{n} R^2_{Z,X_i} \leq R^2_{Z,\{X_1,\ldots,X_n\}} \leq \frac{1}{\gamma_{\emptyset,n}} \sum_{i=1}^{n} R^2_{Z,X_i} \leq \frac{1}{\lambda_{\min}(C)} \sum_{i=1}^{n} R^2_{Z,X_i}$.

Proof. Let the eigenvalues of C^{-1} be $\lambda'_1 \leq \lambda'_2 \leq \ldots \leq \lambda'_n$, with corresponding orthonormal eigenvectors $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}$. We write $\mathbf{b}$ in the basis $\{\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}\}$ as $\mathbf{b} = \sum_i \beta_i \mathbf{e_i}$. Then,

$$R^2_{Z,\{X_1,\ldots,X_n\}} \;=\; \mathbf{b}^T C^{-1} \mathbf{b} \;=\; \sum_i \beta_i^2 \lambda'_i.$$

Because $\lambda'_1 \leq \lambda'_i$ for all i, we get $\lambda'_1 \sum_i \beta_i^2 \leq R^2_{Z,\{X_1,\ldots,X_n\}}$, and $\sum_i \beta_i^2 = \mathbf{b}^T\mathbf{b} = \sum_i R^2_{Z,X_i}$, because the length of the vector $\mathbf{b}$ is independent of the basis it is written in. Also, by definition of the submodularity ratio, $R^2_{Z,\{X_1,\ldots,X_n\}} \leq \frac{\sum_i \beta_i^2}{\gamma_{\emptyset,n}}$. Finally, because $\lambda'_1 = \frac{1}{\lambda_{\max}(C)}$, and using Lemma 2.4, we obtain the result. ■

The next lemma relates the optimal R^2 value using k elements to the optimal R^2 using $k' < k$ elements.

Lemma 3.4 *For each k, let $S^*_k \in \operatorname{argmax}_{|S|\leq k} R^2_{Z,S}$ be an optimal subset of at most k variables. Then, for any $k' = \Theta(k)$ such that $\frac{1}{\lambda_{\min}(C,k)} < k' < k$, we have that $R^2_{Z,S^*_{k'}} \geq R^2_{Z,S^*_k} \cdot \Theta((\frac{k'}{k})^{1/\lambda_{\min}(C,k)})$, for large enough k. In particular, $R^2_{Z,S^*_{k/2}} \geq R^2_{Z,S^*_k} \cdot \Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)})$, for large enough k.*

[4]There is some inconsistency in the literature about naming of greedy algorithms. Forward Regression is sometimes also referred to as Orthogonal Matching Pursuit (OMP). We choose the nomenclature consistent with [10] and [14].

Proof. We first prove that $R^2_{Z,S^*_{k-1}} \geq (1 - \frac{1}{k\lambda_{\min}(C,k)})R^2_{Z,S^*_k}$. Let $T = \mathrm{Res}(Z, S^*_k)$; then, $\mathrm{Cov}(X_i, T) = 0$ for all $X_i \in S^*_k$, and $Z = T + \sum_{X_i \in S^*_k} \alpha_i X_i$, where $\alpha = (\alpha_i) = C^{-1}_{S^*_k} \cdot \mathbf{b}_{S^*_k}$ are the optimal regression coefficients. We write $Z' = Z - T$. For any $X_j \in S^*_k$, by definition of R^2, we have that

$$R^2_{Z',S^*_k\setminus\{X_j\}} \;=\; 1 - \frac{\alpha_j^2 \mathrm{Var}(X_j)}{\mathrm{Var}(Z')} \;=\; 1 - \frac{\alpha_j^2}{\mathrm{Var}(Z')};$$

in particular, this implies that $R^2_{Z',S^*_{k-1}} \geq 1 - \frac{\alpha_j^2}{\mathrm{Var}(Z')}$ for all $X_j \in S^*_k$.

Focus now on j minimizing α_j^2, so that $\alpha_j^2 \leq \frac{\|\alpha\|_2^2}{k}$. As in the proof of Lemma 3.3, by writing α in terms of an orthonormal eigenbasis of $C_{S^*_k}$, one can show that $|\alpha^T C_{S^*_k}\alpha| \geq \|\alpha\|_2^2 \lambda_{\min}(C_{S^*_k})$, or $\|\alpha\|_2^2 \leq \frac{|\alpha^T C_{S^*_k}\alpha|}{\lambda_{\min}(C_{S^*_k})}$. Furthermore, $\alpha^T C_{S^*_k}\alpha = \mathrm{Var}(\sum_{X_i \in S^*_k} \alpha_i X_i) = \mathrm{Var}(Z')$, so $R^2_{Z',S^*_{k-1}} \geq 1 - \frac{1}{k\lambda_{\min}(C_{S^*_k})}$. Finally, by definition, $R^2_{Z',S^*_k} = 1$, so

$$\frac{R^2_{Z,S^*_{k-1}}}{R^2_{Z,S^*_k}} \;\geq\; \frac{R^2_{Z',S^*_{k-1}}}{R^2_{Z',S^*_k}} \;\geq\; 1 - \frac{1}{k\lambda_{\min}(C_{S^*_k})} \;\geq\; 1 - \frac{1}{k\lambda_{\min}(C,k)}.$$

Now, applying this inequality repeatedly, we get

$$R^2_{Z,S^*_{k'}} \geq R^2_{Z,S^*_k} \cdot \prod_{i=k'+1}^{k} (1 - \frac{1}{i\lambda_{\min}(C,i)}).$$

Let $t = \lceil 1/\lambda_{\min}(C,k) \rceil$, so that the previous bound implies $R^2_{Z,S^*_{k'}} \geq R^2_{Z,S^*_k} \cdot \prod_{i=k'+1}^{k} \frac{i-t}{i}$. Most of the terms in the product telescope, giving us a bound of $R^2_{Z,S^*_k} \cdot \prod_{i=1}^{t} \frac{k'-t+i}{k-t+i}$. Since $\prod_{i=1}^{t} \frac{k'-t+i}{k-t+i}$ converges to $(\frac{k'}{k})^t$ with increasing k (keeping t constant), we get that for large k,

$$R^2_{Z,S^*_{k'}} \;\geq\; R^2_{Z,S^*_k} \cdot \Theta((\frac{k'}{k})^t) \;\geq\; R^2_{Z,S^*_k} \cdot \Theta((\frac{k'}{k})^{1/\lambda_{\min}(C,k)}).$$

■

Using the above lemmas, we now prove the main theorem.

Proof of Theorem 3.2. We begin by proving the first inequality. Let S^*_k be the optimum set of variables. Let S^G_i be the set of variables chosen by Forward Regression in the first i iterations, and $S_i = S^*_k \setminus S^G_i$. By monotonicity of R^2 and the fact that $S_i \cup S^G_i \supseteq S^*_k$, we have that $R^2_{Z,S_i \cup S^G_i} \geq \mathrm{OPT}$.

For each $X_j \in S_i$, let $X'_j = \mathrm{Res}(X_j, S^G_i)$ be the residual of X_j conditioned on S^G_i, and write $S'_i = \{X'_j \mid X_j \in S\}$.

We will show that at least one of the X'_i is a good candidate in iteration $i+1$ of Forward Regression. First, the joint contribution of S'_i must be fairly large: $R^2_{Z,\mathrm{Res}(S'_i,S^G_i)} = R^2_{Z,S'_i} \geq \mathrm{OPT} - R^2_{Z,S^G_i}$. Using Definition 2.3, as well as $S^G_i \subseteq S^{\mathrm{FR}}$ and $|S_i| \leq k$,

$$\sum_{X'_j \in S'_i} R^2_{Z,X'_j} \;\geq\; \gamma_{S^G_i,|S_i|} \cdot R^2_{Z,S'_i} \;\geq\; \gamma_{S^{\mathrm{FR}},k} \cdot R^2_{Z,S'_i}.$$

Let ℓ maximize R^2_{Z,X'_ℓ}, i.e., $\ell \in \operatorname{argmax}_{(j:X'_j \in S'_i)} R^2_{Z,X'_j}$. Then we get that

$$R^2_{Z,X'_\ell} \;\geq\; \frac{\gamma_{S^{\text{FR}},k}}{|S'_i|} \cdot R^2_{Z,S'_i} \;\geq\; \frac{\gamma_{S^{\text{FR}},k}}{k} \cdot R^2_{Z,S'_i}.$$

Define $A(i) = R^2_{Z,S^G_i} - R^2_{Z,S^{i-1}_G}$ to be the gain obtained from the variable chosen by Forward Regression in iteration i. Then $R^2_{Z,S^{\text{FR}}} = \sum_{i=1}^k A(i)$. Since the X'_ℓ above was a candidate to be chosen in iteration $i+1$, and Forward Regression chose a variable X_m such that $R^2_{Z,\text{Res}(X_m,S^G_i)} \geq R^2_{Z,\text{Res}(X,S^G_i)}$ for all $X \notin S^G_i$, we obtain that

$$A(i+1) \;\geq\; \frac{\gamma_{S^{\text{FR}},k}}{k} \cdot R^2_{Z,S'_i} \;\geq\; \frac{\gamma_{S^{\text{FR}},k}}{k}(\text{OPT} - R^2_{Z,S^G_i}) \;\geq\; \frac{\gamma_{S^{\text{FR}},k}}{k}(\text{OPT} - \sum_{j=1}^{i} A(j)).$$

Since the above inequality holds for each iteration $i = 1, 2, \ldots, k$, a simple inductive proof establishes the bound $\text{OPT} - \sum_{i=1}^k A(i) \leq \text{OPT} \cdot (1 - \frac{\gamma_{S^{\text{FR}},k}}{k})^k$. Hence,

$$R^2_{Z,S^{\text{FR}}} \;=\; \sum_{i=1}^{k} A(i) \;\geq\; \text{OPT} - \text{OPT}(1 - \frac{\gamma_{S^{\text{FR}},k}}{k})^k \;\geq\; \text{OPT} \cdot (1 - e^{-\gamma_{S^{\text{FR}},k}}).$$

The second inequality follows directly from Lemma 2.4, and the fact that $|S^{\text{FR}}| = k$. By applying the above result after $k/2$ iterations, we obtain $R^2_{Z,S^G_{k/2}} \geq (1 - e^{-\lambda_{\min}(C,k)}) \cdot R^2_{Z,S^*_{k/2}}$. Now, using Lemma 3.4 and monotonicity of R^2, we get

$$R^2_{Z,S^G_k} \;\geq\; R^2_{Z,S^G_{k/2}} \;\geq\; (1 - e^{-\lambda_{\min}(C,k)}) \cdot \Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)}) \cdot R^2_{Z,S^*_k},$$

proving the third inequality. ■

3.2 Orthogonal Matching Pursuit

The second greedy algorithm we analyze is Orthogonal Matching Pursuit (OMP), frequently used in signal processing domains.

Definition 3.5 (Orthogonal Matching Pursuit (OMP)) *The* Orthogonal Matching Pursuit *algorithm for subset selection selects a set S of size k iteratively as follows:*

1: *Initialize $S_0 = \emptyset$.*
2: ***for** each iteration $i + 1$ **do***
3: *Let X_m be a variable maximizing $|\text{Cov}(\text{Res}(Z, S_i), X_m)|$, and set $S_{i+1} = S_i \cup \{X_m\}$.*
4: *Output S_k.*

By applying similar techniques as in the previous section, we can also obtain approximation bounds for OMP. We start by proving the following lemma that lower-bounds the variance of the residual of a variable.

Lemma 3.6 *Let A be the $(n+1)\times(n+1)$ covariance matrix of the normalized variables $Z, X_1, X_2, \ldots, X_n$. Then* $\text{Var}(\text{Res}(Z, \{X_1, \ldots, X_n\})) \geq \lambda_{\min}(A)$.

Proof. The matrix A is of the form $A = \begin{pmatrix} 1 & \mathbf{b}^T \\ \mathbf{b} & C \end{pmatrix}$. We use $A[i,j]$ to denote the matrix obtained by removing the i^{th} row and j^{th} column of A, and similarly for C. Recalling that the (i,j) entry of C^{-1} is $\frac{(-1)^{i+j}\det(C[i,j])}{\det(C)}$, and developing the determinant of A by the first row and column, we can write

$$\begin{aligned}
\det(A) &= \sum_{j=1}^{n+1}(-1)^{1+j}a_{1,j}\det(A[1,j]) \\
&= \det(C) + \sum_{j=1}^{n}(-1)^{j}b_j\det(A[1,j+1]) \\
&= det(C) + \sum_{j=1}^{n}(-1)^{j}b_j\sum_{i=1}^{n}(-1)^{i+1}b_i\det(C[i,j]) \\
&= \det(C) - \sum_{j=1}^{n}\sum_{i=1}^{n}(-1)^{i+j}b_ib_j\det(C[i,j]) \\
&= \det(C)(1-\mathbf{b}^TC^{-1}\mathbf{b}).
\end{aligned}$$

Therefore, using that $\mathrm{Var}(Z) = 1$,

$$\mathrm{Var}(\mathrm{Res}(Z,\{X_1,\ldots,X_n\})) \;=\; \mathrm{Var}(Z) - \mathbf{b}^TC^{-1}\mathbf{b} \;=\; \frac{\det(A)}{\det(C)}.$$

Because $\det(A) = \prod_{i=1}^{n+1}\lambda_i^A$ and $\det(C) = \prod_{i=1}^{n}\lambda_i^C$, and $\lambda_1^A \le \lambda_1^C \le \lambda_2^A \le \lambda_2^C \le \ldots \le \lambda_{n+1}^A$ by the eigenvalue interlacing theorem, we get that $\frac{\det(A)}{\det(C)} \ge \lambda_1^A$, proving the lemma. ■

The above lemma, along with an analysis similar to the proof of Theorem 3.2, can be used to prove the following approximation bounds for OMP:

Theorem 3.7 *The set S^{OMP} selected by orthogonal matching pursuit has the following approximation guarantees:*

$$\begin{aligned}
R^2_{Z,S^{OMP}} &\ge (1 - e^{-(\gamma_{S^{OMP},k}\cdot\lambda_{\min}(C,2k))})\cdot OPT \\
&\ge (1 - e^{-\lambda_{\min}(C,2k)^2})\cdot OPT \\
&\ge (1 - e^{-\lambda_{\min}(C,k)^2})\cdot\Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)})\cdot OPT.
\end{aligned}$$

Proof. We begin by proving the first inequality. Using notation similar to that in the proof of Theorem 3.2, we let S_k^* be the optimum set of k variables, S_i^G the set of variables chosen by OMP in the first i iterations, and $S_i = S_k^* \setminus S_i^G$. For each $X_j \in S_i$, let $X_j' = \mathrm{Res}(X_j, S_i^G)$ be the residual of X_j conditioned on S_i^G, and write $S_i' = \{X_j' \mid X_j \in S\}$.

Consider some iteration $i+1$ of OMP. We will show that at least one of the X_i' is a good candidate in this iteration. Let ℓ maximize $R^2_{Z,X_\ell'}$, i.e., $\ell \in \mathrm{argmax}_{(j:X_j'\in S_i')} R^2_{Z,X_j'}$. By Lemma 3.7,

$$\mathrm{Var}(X_\ell') \;\ge\; \lambda_{\min}(C_{S_G^i\cup\{X_\ell'\}}) \;\ge\; \lambda_{\min}(C,2k).$$

The OMP algorithm chooses a variable X_m to add which maximizes $|\mathrm{Cov}(\mathrm{Res}(Z, S_G^i), X_m)|$. Thus, X_m maximizes

$$\mathrm{Cov}(\mathrm{Res}(Z, S_G^i), X_m)^2 \;=\; \mathrm{Cov}(Z, \mathrm{Res}(X_m, S_G^i))^2 \;=\; R^2_{Z,\mathrm{Res}(X_m,S_G^i)} \cdot \mathrm{Var}(\mathrm{Res}(X_m, S_G^i)).$$

In particular, this implies

$$R^2_{Z,\mathrm{Res}(X_m,S_G^i)} \;\geq\; R^2_{Z,X'_\ell} \cdot \frac{\mathrm{Var}(X'_\ell)}{\mathrm{Var}(\mathrm{Res}(X_m, S_G^i))} \;\geq\; R^2_{Z,X'_\ell} \cdot \frac{\lambda_{\min}(C, 2k)}{\mathrm{Var}(\mathrm{Res}(X_m, S_G^i))} \;\geq\; R^2_{Z,X'_\ell} \cdot \lambda_{\min}(C, 2k),$$

because $\mathrm{Var}(\mathrm{Res}(X_m, S_G^i)) \leq 1$. As in the proof of Theorem 3.2, $R^2_{Z,X'_\ell} \geq \frac{\gamma_{S^{\mathrm{OMP}},k}}{k} \cdot R^2_{Z,S'_i}$, so $R^2_{Z,\mathrm{Res}(X_m,S_G^i)} \geq R^2_{Z,S'_i} \cdot \frac{\lambda_{\min}(C,2k)\cdot\gamma_{S^{\mathrm{OMP}},k}}{k}$. With the same definition of $A(i)$ as in the previous proof, we get that $A(i+1) \geq \frac{\lambda_{\min}(C,2k)\gamma_{S^{\mathrm{OMP}},k}}{k}(P - \sum_{j=1}^{i} A(j))$. An inductive proof now shows that

$$R^2_{Z,S_G} \;=\; \sum_{i=1}^{k} A(i) \;\geq\; (1 - e^{-\lambda_{\min}(C,2k)\cdot\gamma_{S^{\mathrm{OMP}},k}}) \cdot R^2_{Z,S_k^*}.$$

The proofs of the other two inequalities follow the same pattern as the proof for Forward Regression. ■

3.3 Oblivious Algorithm

As a baseline, we also consider a greedy algorithm which completely ignores C and simply selects the k variables individually most correlated with Z.

Definition 3.8 (Oblivious) *The* Oblivious *algorithm for subset selection is as follows: Select the k variables X_i with the largest b_i values.*

Lemma 3.3 immediately implies a simple bound for the Oblivious algorithm:

Theorem 3.9 *The set S^{OBL} selected by the Oblivious algorithm has the following approximation guarantees:*

$$R^2_{Z,S^{OBL}} \;\geq\; \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)} \cdot OPT \;\geq\; \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)} \cdot OPT.$$

Proof. Let S be the set chosen by the Oblivious algorithm, and S_k^* the optimum set of k variables. By definition of the Oblivious algorithm, $\sum_{i\in S} R^2_{Z,X_i} \geq \sum_{i\in S_k^*} R^2_{Z,X_i}$, so using Lemma 3.3, we obtain that

$$R^2_{Z,S} \geq \frac{\sum_{i\in S} R^2_{Z,X_i}}{\lambda_{\max}(C,k)} \;\geq\; \frac{\sum_{i\in S_k^*} R^2_{Z,X_i}}{\lambda_{\max}(C,k)} \;\geq\; \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)} R^2_{Z,S_k^*}.$$

The second inequality of the theorem follows directly from Lemma 2.4. ■

4 Dictionary Selection Bounds

To demonstrate the wider applicability of the approximate submodularity framework, we next obtain a tighter analysis for two greedy algorithms for the dictionary selection problem, introduced in [8].

4.1 The Algorithm $\mathrm{SDS_{MA}}$

The $\mathrm{SDS_{MA}}$ algorithm generalizes the Oblivious greedy algorithm to the problem of dictionary selection. It replaces the $R^2_{Z_j,S}$ term in Definition 2.2 with its modular approximation $f(Z_j, S) = \sum_{i \in S} R^2_{Z_j,X_i}$. Thus, it greedily tries to maximize the function $\hat{F}(D) = \sum_{j=1}^{s} \max_{S \subset D, |S|=k} f(Z_j, S)$, over sets D of size at most d; the inner maximum can be computed efficiently using the Oblivious algorithm.

Definition 4.1 ($\mathrm{SDS_{MA}}$) *The $\mathrm{SDS_{MA}}$ algorithm for dictionary selection selects a dictionary D of size d iteratively as follows:*

1: *Initialize $D_0 = \emptyset$.*
2: ***for** each iteration $i + 1$ **do***
3: *Let X_m be a variable maximizing $\hat{F}(D \cup \{X_m\})$, and set $S_{i+1} = S_i \cup \{X_m\}$.*
4: *Output D_d.*

Using Lemma 3.3, we can obtain the following multiplicative approximation guarantee for $\mathrm{SDS_{MA}}$:

Theorem 4.2 *Let D^{MA} be the dictionary selected by the $\mathrm{SDS_{MA}}$ algorithm, and D^* the optimum dictionary of size $|D| \leq d$, with respect to the objective $F(D)$ from Definition 2.2. Then,*

$$F(D^{MA}) \;\geq\; \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)}(1 - \frac{1}{e}) \cdot F(D^*) \;\geq\; \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)}(1 - \frac{1}{e}) \cdot F(D^*).$$

Proof. Let $\hat{D}$ be a dictionary of size d maximizing $\hat{F}(D)$. Because $f(Z_j, S)$ is monotone and modular in S, $\hat{F}$ is a monotone, submodular function. Hence, using the submodularity results of Nemhauser et al. [12] and the optimality of $\hat{D}$ for $\hat{F}$,

$$\hat{F}(D^{\mathrm{MA}}) \;\geq\; \hat{F}(\hat{D})(1 - \frac{1}{e}) \;\geq\; \hat{F}(D^*)(1 - \frac{1}{e}).$$

Now, by applying Lemma 3.3 for each Z_j, it is easy to show that $\hat{F}(D^*) \geq \gamma_{\emptyset,k} \cdot F(D^*)$, and similarly $\hat{F}(D^{\mathrm{MA}}) \leq \lambda_{\max}(C,k) \cdot F(D^{\mathrm{MA}})$. Thus we get $F(D^{\mathrm{MA}}) \geq \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)}(1 - \frac{1}{e})F(D^*)$.

The second part now follows from Lemma 2.4. ■

Note that these bounds significantly improve the previous additive approximation guarantee obtained in [8]: $F(D^{\mathrm{MA}}) \geq (1 - \frac{1}{e})F(D^*) - (2 - \frac{1}{e})k \cdot \mu(C)$. In particular, when $\mu(C) > \Theta(1/k)$, i.e., even just one pair of variables has moderate correlation, the approximation guarantee of Krause and Cevher becomes trivial.

4.2 The Algorithm $\mathrm{SDS_{OMP}}$

We also obtain a multiplicative approximation guarantee for the greedy $\mathrm{SDS_{OMP}}$ algorithm, introduced by Krause and Cevher for dictionary selection. Our bounds for $\mathrm{SDS_{OMP}}$ are much stronger than the additive bounds obtained by Krause and Cevher. However, for both our results and theirs, the performance guarantees for $\mathrm{SDS_{OMP}}$ are much weaker than those for $\mathrm{SDS_{MA}}$.

The $\mathrm{SDS_{OMP}}$ algorithm generalizes the Orthogonal Matching Pursuit algorithm for subset selection to the problem of dictionary selection. In each iteration, it adds a new element to the currently selected dictionary by using Orthogonal Matching Pursuit to approximate the estimation of $\max_{|S|=k} R^2_{Z_j,S}$.

Definition 4.3 ($\mathrm{SDS_{OMP}}$) *The $\mathrm{SDS_{OMP}}$ algorithm for dictionary selection selects a dictionary D of size d iteratively as follows:*

1: *Initialize* $D_0 = \emptyset$.
2: ***for*** *each iteration* $i+1$ ***do***
3: *Let* X_m *be a variable maximizing* $\sum_{j=1}^{s} R^2_{Z_j,S_{OMP}(D_i \cup \{X_m\},Z_j,k)}$ *where* $S_{OMP}(D,Z,k)$ *denotes the set selected by Orthogonal Matching Pursuit for predicting* Z *using* k *variables from* D.
4: *Set* $S_{i+1} = S_i \cup \{X_m\}$.
5: *Output* D_d.

We now show how to obtain a multiplicative approximation guarantee for $\mathrm{SDS}_{\mathrm{OMP}}$. The following definitions are key to our analysis; the first two are from Definition 2.2 and Theorem 4.2.

$$F(D) = \sum_{j=1}^{s} \max_{S \subset D, |S|=k} R^2_{Z_j,S},$$

$$\hat{F}(D) = \sum_{j=1}^{s} \max_{S \subset D, |S|=k} f(Z_j, S),$$

$$\tilde{F}(D) = \sum_{j=1}^{s} R^2_{Z_j,S_{\mathrm{OMP}}(D,Z_j,k)}.$$

We first prove the following lemma about approximating the function $\hat{F}(D)$ by $\tilde{F}(D)$:

Lemma 4.4 *For any set* D*, we have that*

$$\frac{(1-e^{-\lambda_{\min}(C,2k)^2})}{\lambda_{\max}(C,k)} \cdot \hat{F}(D) \;\; \leq \;\; \tilde{F}(D) \;\; \leq \;\; \frac{\hat{F}(D)}{\gamma_{\emptyset,k}}.$$

Proof. Using Theorem 3.7 and Lemma 3.3 and summing up over all the Z_j terms, we obtain that

$$\tilde{F}(D) \;\; \geq \;\; (1 - e^{-\lambda_{\min}(C,2k)^2}) \cdot F(D) \;\; \geq \;\; (1 - e^{-\lambda_{\min}(C,2k)^2}) \frac{\hat{F}(D)}{\lambda_{\max}(C,k)}.$$

Similarly, using Lemma 3.3 and the fact that $\max_{S \subset D, |S|=k} R^2_{Z_j,S} \geq R^2_{Z_j,S_{OMP}(D,Z_j,k)}$, we have

$$\hat{F}(D) \;\; \geq \;\; \gamma_{\emptyset,k} \cdot F(D) \;\; \geq \;\; \gamma_{\emptyset,k} \cdot \tilde{F}(D).$$

■

Using the above lemma, we now prove the following bound for $\mathrm{SDS}_{\mathrm{OMP}}$:

Theorem 4.5 *Let* D^{OMP} *be the dictionary selected by the* $\mathrm{SDS}_{\mathrm{OMP}}$ *algorithm, and* D^* *the optimum dictionary of size* $|D| \leq d$*, with respect to the objective* $F(D)$ *from Definition 2.2. Then,*

$$F(D^{OMP}) \;\geq\; F(D^*) \cdot \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)} \cdot \frac{(1-e^{-(p \cdot \gamma_{\emptyset,k})})}{d - d \cdot p \cdot \gamma_{\emptyset,k} + 1} \;\geq\; F(D^*) \cdot \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)} \cdot \frac{(1-e^{-(p \cdot \gamma_{\emptyset,k})})}{d - d \cdot p \cdot \gamma_{\emptyset,k} + 1},$$

where $p = \frac{1}{\lambda_{\max}(C,k)} \cdot (1 - e^{-\lambda_{\min}(C,2k)^2})$.

Proof. Let $\hat{D}$ be the dictionary of size d that maximizes $\hat{F}(D)$. We first prove that $\hat{F}(D^{\text{OMP}})$ is a good approximation to $\hat{F}(\hat{D})$.

Let S_i^G be the variables chosen by SDS_{OMP} after i iterations. Define $S_i = \hat{D} \setminus S_i^G$. By monotonicity of $\hat{F}$, we have that $\hat{F}(S_i \cup S_i^G) \geq \hat{F}(\hat{D})$.

Let $\hat{X} \in S_i$ be the variable maximizing $\hat{F}(S_i^G \cup \{\hat{X}\})$, and similarly $\tilde{X} \in S_i$ be the variable maximizing $\tilde{F}(S_i^G \cup \{\tilde{X}\})$.

Since $\hat{F}$ is a submodular function, it is easy to show (using an argument similar to the proof of Theorem 3.2) that $\hat{F}(S_i^G \cup \{\hat{X}\}) - \hat{F}(S_i^G) \geq \frac{\hat{F}(\hat{D}) - \hat{F}(S_i^G)}{d}$.

Now, using Lemma 4.4 above, and the optimality of $\tilde{X}$ for $\tilde{F}(S_i^G \cup \{\tilde{X}\})$, we obtain that

$$\frac{1}{\gamma_{\emptyset,k}} \cdot \hat{F}(S_i^G \cup \{\tilde{X}\}) \quad \geq \quad \tilde{F}(S_i^G \cup \{\tilde{X}\}) \quad \geq \quad \tilde{F}(S_i^G \cup \{\hat{X}\}) \quad \geq \quad p \cdot \hat{F}(S_i^G \cup \{\hat{X}\}).$$

Thus, $\hat{F}(S_i^G \cup \{\tilde{X}\}) \geq p \cdot \gamma_{\emptyset,k} \cdot \hat{F}(S_i^G \cup \{\hat{X}\})$, or

$$\hat{F}(S_i^G \cup \{\tilde{X}\}) - \hat{F}(S_i^G) \geq p \cdot \gamma_{\emptyset,k} \cdot (\hat{F}(S_i^G \cup \{\hat{X}\}) - \hat{F}(S_i^G)) - (1 - p \cdot \gamma_{\emptyset,k})\hat{F}(S_i^G).$$

Define $A(i) = \hat{F}(S_i^G) - \hat{F}(S_{i-1}^G)$ to be the gain, with respect to $\hat{F}$, obtained from the variable chosen by SDS_{OMP} in iteration i. Then $\hat{F}(D^{\text{OMP}}) = \sum_{i=1}^{d} A(i)$. From the preceding paragraphs, we obtain

$$A(i+1) \geq \frac{p \cdot \gamma_{\emptyset,k}}{d} \cdot (\hat{F}(\hat{D}) - (1 + \frac{d}{p \cdot \gamma_{\emptyset,k}} - d) \sum_{j=1}^{i} A(j)).$$

Since the above inequality holds for each iteration $i = 1, 2, \ldots, d$, a simple inductive proof shows that

$$\hat{F}(\hat{D}) - \sum_{i=1}^{d} A(i) \leq \hat{F}(\hat{D}) \cdot (1 - \frac{p\gamma_{\emptyset,k}}{d})^d + (d - dp\gamma_{\emptyset,k}) \cdot \sum_{i=1}^{d} A(i).$$

Rearranging the terms and simplifying, we get that

$$\hat{F}(D^{\text{OMP}}) \quad = \quad \textstyle\sum_{i=1}^{d} A(i) \quad \geq \quad \hat{F}(\hat{D}) \cdot \frac{(1 - e^{-(p \cdot \gamma_{\emptyset,k})})}{d - dp\gamma_{\emptyset,k} + 1} \quad \geq \quad \hat{F}(D^*) \cdot \frac{(1 - e^{-(p \cdot \gamma_{\emptyset,k})})}{d - dp\gamma_{\emptyset,k} + 1},$$

where the last inequality is due to the optimality of $\hat{D}$ for $\hat{F}$.

Now, using Lemma 3.3 for each Z_j term, it can be easily seen that $\hat{F}(D^*) \geq \gamma_{\emptyset,k} \cdot F(D^*)$. Similarly, using Lemma 3.3 on the set D^{OMP}, we have $F(D^{\text{OMP}}) \geq \frac{1}{\lambda_{\max}(C,k)} \cdot \hat{F}(D^{\text{OMP}})$.

Using the above inequalities, we therefore get the desired bound

$$F(D^{\text{OMP}}) \geq F(D^*) \cdot \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)} \cdot \frac{(1 - e^{-(p \cdot \gamma_{\emptyset,k})})}{d - d \cdot p \cdot \gamma_{\emptyset,k} + 1}.$$

The second inequality of the Theorem now follows directly from Lemma 2.4. ■

5 Experiments

In this section, we evaluate Forward Regression (FR) and OMP empirically, on two real-world and one synthetic data set. We compare the two algorithms against an optimal solution (OPT), computed using exhaustive search, the Oblivious greedy algorithm (OBL), and the L1-regularization/Lasso (L1) algorithm (in the implementation of Koh et al. [7]). Beyond the algorithms' performance, we also compute the various spectral parameters from which we can derive lower bounds. Specifically, these are

1. the submodularity ratio: $\gamma_{S^{\text{FR}},k}$, where S^{FR} is the subset selected by forward regression.

2. the smallest sparse eigenvalues $\lambda_{\min}(C,k)$ and $\lambda_{\min}(C,2k)$. (In some cases, computing $\lambda_{\min}(C,2k)$ was not computationally feasible due to the problem size.)

3. the sparse inverse condition number $\kappa(C,k)^{-1}$. As mentioned earlier, the sparse inverse condition number $\kappa(C,k)$ is strongly related to the Restricted Isometry Property in [1].

4. the smallest eigenvalue $\lambda_{\min}(C) = \lambda_{\min}(C,n)$ of the entire covariance matrix.

The aim of our experiments is twofold: First, we wish to evaluate which among the submodular and spectral parameters are good predictors of the performance of greedy algorithms in practice. Second, we wish to highlight how the theoretical bounds for subset selection algorithms reflect on their actual performance. Our analytical results predict that Forward Regression should outperform OMP, which in turn outperforms Oblivious. For Lasso, it is not known whether strong multiplicative bounds, like the ones we proved for Forward Regression or OMP, can be obtained.

5.1 Data Sets

Because several of the spectral parameters (as well as the optimum solution) are **NP**-hard to compute, we restrict our experiments to data sets with $n \leq 30$ features, from which $k \leq 8$ are to be selected. We stress that the greedy algorithms themselves are very efficient, and the restriction on data set sizes is only intended to allow for an adequate evaluation of the results.

Each data set contains $m > n$ samples, from which we compute the empirical covariance matrix (analogous to the Gram matrix in sparse approximation) between all observation variables and the predictor variable; we then normalize it to obtain C and $\mathbf{b}$. We evaluate the performance of all algorithms in terms of their R^2 fit; thus, we implicitly treat C and $\mathbf{b}$ as the ground truth, and also do not separate the data sets into training and test cases.

Our data sets are the *Boston Housing Data*, a data set of *World Bank Development Indicators*, and a synthetic data set generated from a distribution similar to the one used by Zhang [17]. The *Boston Housing Data* (available from the UCI Machine Learning Repository) is a small data set frequently used to evaluate ML algorithms. It comprises $n = 15$ features (such as crime rate, property tax rates, etc.) and $m = 516$ observations. Our goal is to predict housing prices from these features. The *World Bank Data* (available from `http://databank.worldbank.org`) contains an extensive list of socio-economic and health indicators of development, for many countries and over several years. We choose a subset of $n = 29$ indicators for the years 2005 and 2006, such that the values for all of the $m = 65$ countries are known for each indicator. (The data set does not contain all indicators for each country.) We choose to predict the average life expectancy for those countries.

To perform tests in a controlled fashion, we also generate random instances from a known distribution similar to [17]: There are $n = 29$ features, and $m = 100$ data points are generated from a joint Gaussian distribution with moderately high correlations of 0.6. The target vector is obtained by generating coefficients uniformly from 0 to 10 along each dimension, and adding noise with variance $\sigma^2 = 0.1$. Notice that the target vector is not truly sparse. The plots we show are the average R^2 values for 20 independent runs of the experiment.

5.2 Results

We run the different subset selection algorithms for values of k from 2 through 8, and plot the R^2 values for the selected sets. Figures 1, 3 and 5 show the results for the three data sets. The main insight is that on all data sets, Forward Regression performs optimally or near-optimally, and OMP is only slightly worse. Lasso performs somewhat worse on all data sets, and, not surprisingly, the baseline Oblivious algorithm performs even worse. The order of performance of the greedy algorithms match the order of the strength of the theoretical bounds we derived for them.

On the World Bank data (Figure 3), all algorithms perform quite well with just 2–3 features already. The main reason is that adolescent birth rate is by itself highly predictive of life expectancy, so the first feature selected by all algorithms already contributes high R^2 value.

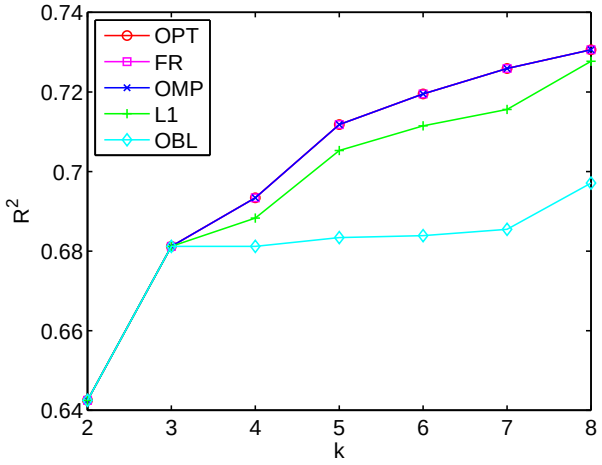

Figure 1: Boston Housing R^2

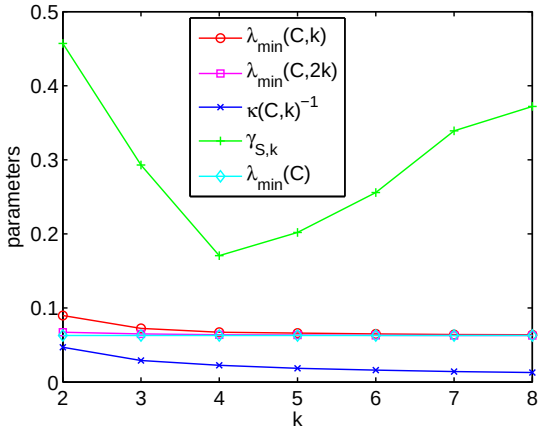

Figure 2: Boston Housing parameters

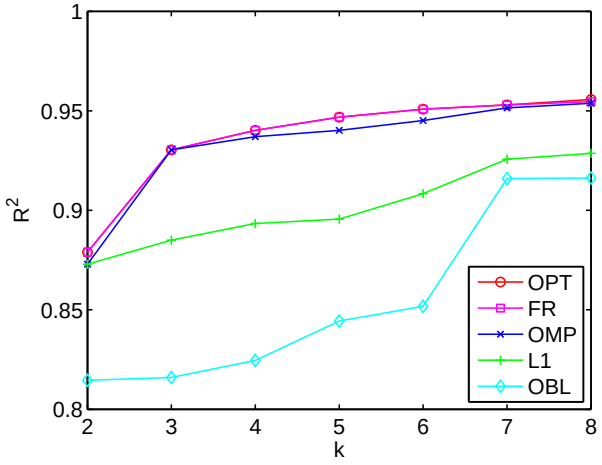

Figure 3: World Bank R^2

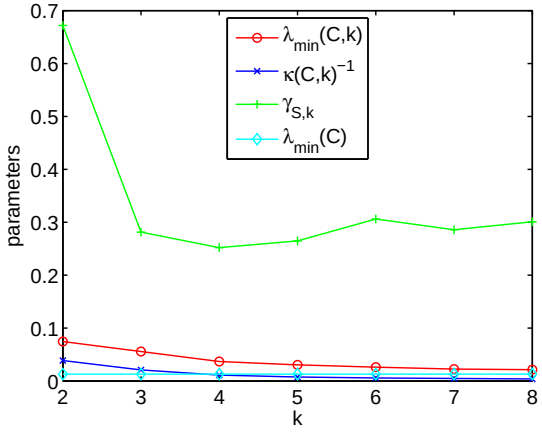

Figure 4: World Bank parameters

Figures 2, 4 and 6 show the different spectral quantities for the data sets, for varying values of k. Both of the real-world data sets are nearly singular, as evidenced by the small $\lambda_{\min}(C)$ values. In fact, the near

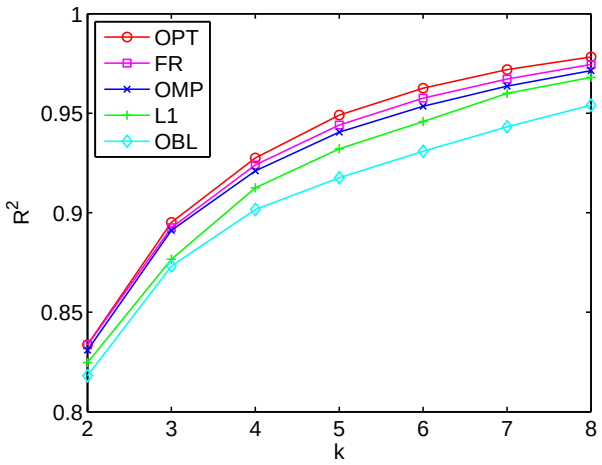

Figure 5: Synthetic Data R^2

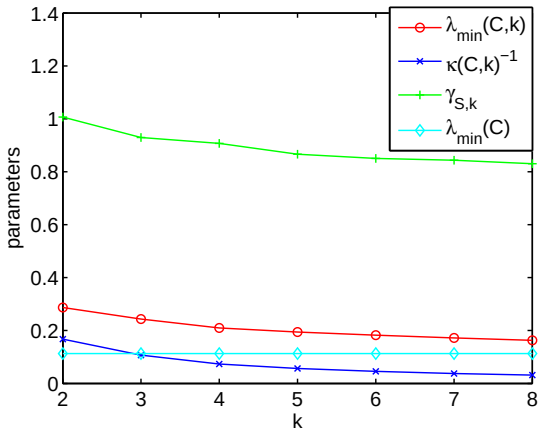

Figure 6: Synthetic Data parameters

singularities manifest themselves for small values of k already; in particular, since $\lambda_{\min}(C, 2)$ is already small, we observe that there are pairs of highly correlated observations variables in the data sets. Thus, the bounds on approximation we would obtain by considering merely $\lambda_{\min}(C, k)$ or $\lambda_{\min}(C, 2k)$ would be quite weak. Notice, however, that they are still quite a bit stronger than the inverse condition number $\kappa(C, k)^{-1}$: this bound — which is closely related to the RIP property frequently at the center of sparse approximation analysis — takes on much smaller values, and thus would be an even looser bound than the eigenvalues.

On the other hand, the submodularity ratios $\gamma_{S^{\text{FR}},k}$ for all the data sets are much larger than the other spectral quantities (almost 5 times larger, on average, than the corresponding $\lambda_{\min}(C)$ values). Notice that unlike the other quantities, the submodularity ratios are not monotonically decreasing in k — this is due to the dependency of $\gamma_{S^{\text{FR}},k}$ on the set S^{FR}, which is different for every k.

The discrepancy between the small values of the eigenvalues and the good performance of all algorithms shows that bounds based solely on eigenvalues can sometimes be loose. Significantly better bounds are obtained from the submodularity ratio $\gamma_{S^{\text{FR}},k}$, which takes on values above 0.2, and significantly larger in many cases. While not entirely sufficient to explain the performance of the greedy algorithms, it shows that the near-singularities of C do not align unfavorably with $\mathbf{b}$, and thus do not provide an opportunity for strong supermodular behavior that adversely affects greedy algorithms.

The synthetic data set we generated is somewhat further from singular, with $\lambda_{\min}(C) \approx 0.11$. However, the same patterns persist: the simple eigenvalue based bounds, while somewhat larger for small k, still do not fully predict the performance of greedy algorithms, whereas the submodularity ratio here is close to 1 for all values of k. This shows that the near-singularities do not at all provide the possibility of strongly supermodular benefits of sets of variables. Indeed, the plot of R^2 values on the synthetic data is concave, an indicator of submodular behavior of the function.

The above observations suggest that bounds based on the submodularity ratio are better predictors of the performance of greedy algorithms, followed by bounds based on the sparse eigenvalues, and finally those based on the condition number or RIP property.

5.3 Narrowing the gap between theory and practice

Our theoretical bounds, though much stronger than previous results, still do not fully predict the observed near-optimal performance of Forward Regression and OMP on the real-world datasets. In particular, for Forward Regression, even though the submodularity ratio is less than 0.4 for most cases, implying a theoretical guarantee of roughly $1 - e^{-0.4} \approx 33\%$, the algorithm still achieves near-optimal performance. While gaps between worst-case bounds and practical performance are commonplace in algorithmic analysis, they also suggest that there is scope for further improving the analysis, by looking at more fine-grained parameters.

Indeed, a slightly more careful analysis of the proof of Theorem 3.2 and our definition of the submodularity ratio reveals that we do not really need to calculate the submodularity ratio over all sets S of size k while analyzing the greedy steps of Forward Regression. We can ignore sets S whose submodularity ratio is low, but whose marginal contribution to the current R^2 is only a small fraction (say, at most ϵ). This is because the proof of Theorem 3.2 shows that for each iteration $i+1$, we only need to consider the submodularity ratio for the set $S_i = S_k^* \setminus S_i^G$, where S_i^G is the set selected by the greedy algorithm after i iterations, and S_k^* is the optimal k-subset. Thus, if $R^2_{Z,S_i \cup S_i^G} \leq (1+\epsilon) \cdot R^2_{Z,S_i^G}$, then the currently selected set must already be within a factor $\frac{1}{1+\epsilon}$ of optimal.

By carefully pruning such sets (using $\epsilon = 0.2$) while calculating the submodularity ratio, we see that the resulting values of $\gamma_{S^{\text{FR}},k}$ are much higher (more than 0.8), thus significantly reducing the gap between the theoretical bounds and experimental results. Table 1 shows the values of $\gamma_{S^{\text{FR}},k}$ obtained using this method.

The results suggest an interesting direction for future work: namely, to characterize for which sets the submodular behavior of R^2 really matters.

Data Set	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Boston	0.9	0.91	1.02	1.21	1.36	1.54	1.74
World Bank	0.8	0.81	0.81	0.81	0.94	1.19	1.40

Table 1: Improved estimates for submodularity ratio

6 Discussion and Concluding Remarks

In this paper, we analyze greedy algorithms using the notion of submodularity ratio, which captures how close to submodular an objective function (in our case the R^2 measure of statistical fit) is. Using submodular analysis, coupled with spectral techniques, we prove the strongest known approximation guarantees for commonly used greedy algorithms for subset selection and dictionary selection. Our bounds help explain why greedy algorithms perform well in practice even in the presence of strongly correlated data, and are substantiated by experiments on real-world and synthetic datasets. The experiments show that the submodularity ratio is a much stronger predictor of the performance of greedy algorithms than previously used spectral parameters. We believe that our techniques for analyzing greedy algorithms using a notion of "approximate submodularity" are not specific to subset selection and dictionary selection, and could also be used to analyze other problems in compressed sensing and sparse recovery.

References

[1] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[2] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *ACM Symposium on Theory of Computing*, 2008.

[3] G. Diekhoff. *Statistics for the Social and Behavioral Sciences*. Wm. C. Brown Publishers, 2002.

[4] D. Donoho. For most large underdetermined systems of linear equations, the minimal 11-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[5] A. Gilbert, S. Muthukrishnan, and M. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[6] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.

[7] K. Koh, S. Kim, and S. Boyd. l1_ls: Simple Matlab Solver for l1-regularized Least Squares Problems, 2008. http://www.stanford.edu/ boyd/l1_ls.

[8] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proc. ICML*, 2010.

[9] A. C. Lozano, G. Swirszcz, and N. Abe. Grouped orthogonal matching pursuit for variable selection and prediction. In *Proc. NIPS*, 2009.

[10] A. Miller. *Subset Selection in Regression*. Chapman and Hall, second edition, 2002.

[11] B. Natarajan. Sparse approximation solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

[12] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[13] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society*, 58:267–288, 1996.

[14] J. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Information Theory*, 50:2231–2242, 2004.

[15] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Information Theory*, 51:1030–1051, 2006.

[16] J. Tropp, A. Gilbert, S. Muthukrishnan, and M. Strauss. Improved sparse approximation over quasi-incoherent dictionaries. In *Proc. IEEE-ICIP*, 2003.

[17] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proc. NIPS*, 2008.

[18] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[19] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2451–2457, 2006.

[20] S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Proc. NIPS*, 2009.

A Estimating $\lambda_{\min}(C,k)$

Several of our approximation guarantees are phrased in terms of $\lambda_{\min}(C,k)$. Finding the exact value of $\lambda_{\min}(C,k)$ is **NP**-hard in general; here, we show how to estimate lower and upper bounds. Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of C, and $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}$ the corresponding eigenvectors. A first simple bound can be obtained directly from the eigenvalue interlacing theorem: $\lambda_1 \leq \lambda_{\min}(C,k) \leq \lambda_{n-k+1}$.

One case in which good lower bounds on $\lambda_{\min}(C,k)$ can possibly be obtained is when only a small (constant) number of the λ_i are small. The following lemma allows a bound in terms of any λ_j; however, since the running time by the implied algorithm is exponential in j, and the quality of the bound depends on λ_j, it is useful only in the special case when $\lambda_j \gg 0$ for a small constant j.

Lemma A.1 *Let V_j be the vector space spanned by the eigenvectors $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_j}$, and define*

$$\beta_j = \max_{\mathbf{y}\in V_j, \mathbf{x}\in\mathbb{R}^n, \|\mathbf{y}\|_2=\|\mathbf{x}\|_2=1, \|\mathbf{x}\|_0\leq k} |\mathbf{x}\cdot\mathbf{y}|.$$

Then, $\lambda_{\min}(C,k) \geq \lambda_{j+1}\cdot(1-\beta_j)$.

Proof. Let $\mathbf{x}' \in \mathbb{R}^n, \|\mathbf{x}'\|_2 = 1, \|\mathbf{x}\|_0 \leq k$ be an eigenvector corresponding to $\lambda_{\min}(C,k)$. Let α_i be the coefficients of the representation of $\mathbf{x}'$ in terms of the $\mathbf{e_i}$: $\mathbf{x}' = \sum_{i=1}^n \alpha_i \mathbf{e_i}$. Thus, $\sum_{i=1}^n \alpha_i^2 = 1$, and we can write

$$\lambda_{\min}(C,k) \;=\; \mathbf{x'}^T C \mathbf{x}' \;=\; \sum_{i=1}^n \alpha_i^2 \lambda_i \;\geq\; \lambda_{j+1}(1-\sum_{i=1}^j \alpha_i^2).$$

Since $\sum_{i=1}^j \alpha_i^2$ is the length of the projection of $\mathbf{x}$ onto V_j, we have

$$\sum_{i=1}^j \alpha_i^2 \;=\; \max_{\mathbf{y}\in V_j, \|\mathbf{y}\|_2=1} |\mathbf{x}'\cdot\mathbf{y}| \;\leq\; \max_{\mathbf{y}\in V_j, \mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_2=\|\mathbf{y}\|_2=1, \|\mathbf{x}\|_0\leq k} |\mathbf{y}\cdot\mathbf{x}|,$$

completing the proof. ■

Since all the λ_j can be computed easily, the crux in using this bound is finding a good bound on β_j. Next, we show a PTAS (Polynomial-Time Approximation Scheme) for approximating β_j, for any constant j.

Lemma A.2 *For every $\epsilon > 0$, there is a $1-\epsilon$ approximation for calculating β_j, running in time $O((\frac{n}{\epsilon})^j)$.*

Proof. Any vector $\mathbf{y} \in V_j$ with $\|\mathbf{y}\|_2 = 1$ can be written as $\mathbf{y} = \sum_{i=1}^j \eta_i \mathbf{e_i}$ with $\eta_i \in [-1,1]$ for all i. The idea of our algorithm is to exhaustively search over all $\mathbf{y}$, as parametrized by their η_i entries. To make the search finite, the entries are discretized to multiples of $\delta = \epsilon \cdot \sqrt{k/(nj)}$. The total number of such vectors to search over is $(2/\delta)^j \leq (n/\epsilon)^j$.

Let $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ attain the maximum in the definition of β_j, and write $\hat{\mathbf{y}} = \sum_{i=1}^j \hat{\eta}_i \mathbf{e_i}$. For each i, let η_i be $\hat{\eta}_i$, rounded to the nearest multiple of δ, and $\mathbf{y} = \sum_{i=1}^j \eta_i \mathbf{e_i}$. Then, $\|\hat{\mathbf{y}} - \mathbf{y}\|_2 \leq \|\delta \sum_{i=1}^j \mathbf{e_j}\|_2 = \delta\sqrt{j}$.

The vector $\mathbf{x}' = \mathrm{argmax}_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_2=1, \|\mathbf{x}\|_0\leq k} |\mathbf{y}\cdot\mathbf{x}|$ is of the following form: Let I be the set of k indices i such that $|y_i|$ is largest, and $\gamma = \sqrt{\sum_{i\in I} y_i^2}$. Then, $x'_i = 0$ for $i \notin I$ and $x'_i = y_i/\gamma$ for $i \in I$. Notice that given $\mathbf{y}$, we can easily find $\mathbf{x}'$, and because $|\hat{\mathbf{x}}\cdot\mathbf{y}| \leq |\mathbf{x}'\cdot\mathbf{y}| \leq |\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}|$, we have

$$\frac{||\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}| - |\mathbf{x}'\cdot\mathbf{y}||}{|\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}|} \;\leq\; \frac{||\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}| - |\hat{\mathbf{x}}\cdot\mathbf{y}||}{|\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}|} \;\leq\; \frac{\|\hat{\mathbf{x}}\|_2\|\hat{\mathbf{y}}-\mathbf{y}\|_2}{|\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}|} \;\leq\; \frac{\delta\sqrt{j}}{|\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}|} \;\leq\; \delta\sqrt{jn/k}.$$

The last inequality follows since the sum of the k largest entries of $\hat{\mathbf{y}}$ is at least $k/\sqrt{n}$, so by setting $x_i = 1/\sqrt{k}$ for each of those coordinates, we can attain at least an inner product of $\sqrt{k/n}$, and the inner product with $\hat{\mathbf{x}}$ cannot be smaller.

The value output by the exhaustive search over all discretized values is at least $|\mathbf{x}' \cdot \mathbf{y}|$, and thus within a factor of $1 - \frac{\delta\sqrt{j}n}{k} = 1 - \epsilon$ of the maximum value, attained by $\hat{\mathbf{x}}, \hat{\mathbf{y}}$. ■