

RESTRICTED STRONG CONVEXITY IMPLIES WEAK SUBMODULARITY¹

BY ETHAN R. ELENBERG*, RAJIV KHANNA*, ALEXANDROS G. DIMAKIS*
AND SAHAND NEGAHBAN[†]

University of Texas at Austin and Yale University[†]*

We connect high-dimensional subset selection and submodular maximization. Our results extend the work of Das and Kempe [In *ICML* (2011) 1057–1064] from the setting of linear regression to arbitrary objective functions. For greedy feature selection, this connection allows us to obtain strong multiplicative performance bounds on several methods without statistical modeling assumptions. We also derive recovery guarantees of this form under standard assumptions. Our work shows that greedy algorithms perform within a constant factor from the best possible subset-selection solution for a broad class of general objective functions. Our methods allow a direct control over the number of obtained features as opposed to regularization parameters that only implicitly control sparsity. Our proof technique uses the concept of *weak submodularity* initially defined by Das and Kempe. We draw a connection between convex analysis and submodular set function theory which may be of independent interest for other statistical learning applications that have combinatorial structure.

1. Introduction. Sparse modeling is central in modern data analysis and high-dimensional statistics since it provides interpretability and robustness. Given a large set of p features, we wish to build a model using only a small subset of k features: the central combinatorial question is how to choose the optimal feature subset. Specifically, we are interested in optimizing over sparse parameter vectors β and consider problems of the form:

$$(1) \quad \bar{\beta}^k \in \arg \max_{\beta: \|\beta\|_0 \leq k} l(\beta),$$

for some function $l(\cdot)$. This is a very general framework: the function $l(\cdot)$ can be a linear regression R^2 objective, a generalized linear model (GLM) likelihood, a graphical model learning objective, or an arbitrary M -estimator [40]. This *subset selection* problem is NP-hard [37] even for the sparse linear regression objective, and a vast body of literature has analyzed different approximation algorithms under various assumptions.

Received July 2017; revised October 2017.

¹Supported by NSF Grants CCF 1344179, 1344364, 1407278, 1422549, 1618689, 1723128, DMS 1723052, IIS 1421729, and ARO YIP W911NF-14-1-0258.

MSC2010 subject classifications. Primary 68W25; secondary 62J02, 68W40.

Key words and phrases. Submodular functions, greedy algorithms, restricted strong convexity, subset selection.

The Restricted Isometry Property (RIP) and the (closely related) Restricted Eigenvalue property are conditions on $l(\beta)$ that allow convex relaxations and greedy algorithms to solve the subset selection problem within provable approximation guarantees. In parallel work, several authors have demonstrated that the subset selection problem can be connected to submodular optimization [2, 20, 48, 49] and that greedy algorithms are widely used for iteratively building good feature sets.

The mathematical connection between submodularity and RIP was made explicit by Das and Kempe [13] for linear regression. Specifically, they showed that when $l(\beta)$ is the R^2 objective, it satisfies a weak form of submodularity when the linear measurements satisfy RIP. Note that for a given set of features S , the function $l(\beta_S)$ with support restricted to S can be thought of as a set function and this is key in this framework. Using this novel concept of *weak submodularity*, they established strong multiplicative bounds on the performance of greedy algorithms for subset selection and dictionary selection. Work by Bach [2] in the linear regression setting discusses the notion of *suppressors*; however, that condition is stronger than the weak submodularity assumption. Krause and Cevher [28] draw similar connections between submodularity and sparse regression, but they require a much stronger coherence-based assumption.

In this paper, we extend this machinery beyond linear regression, to any function $l(\beta)$. To achieve this, we need the proper generalization of the Restricted Eigenvalue and RIP conditions for arbitrary functions. This was obtained by Negahban et al. [40] and is called *Restricted Strong Convexity*. The title of this paper should now be clear: we show that any objective function that satisfies *restricted strong convexity* (and a natural smoothness assumption) of [40] must be weakly submodular.

We establish multiplicative approximation bounds on the performance of greedy algorithms, including (generalized) Orthogonal Matching Pursuit and Forward Stepwise Regression, for general likelihood functions using our connection. To the best of our knowledge, this is the first analysis of connecting a form of submodularity to the objective function's strong concavity and smoothness. Our approach provides sharp approximation bounds in any setting where these fundamental structural properties are well understood, for example, generalized linear models.

Contrary to prior work, we require no assumptions on the sparsity of the underlying problem. Rather, we obtain a deterministic result establishing multiplicative approximation guarantees from the best-case sparse solution. Our results improve over previous work by providing bounds on a solution that is guaranteed to match the desired sparsity. Convex methods such as ℓ_1 regularized objectives require strong assumptions on the model, such as the irrepresentability conditions on the feature vectors, in order to provide exact sparsity guarantees on the recovered solution.

Our main result is that for any function $l(\cdot)$ that satisfies M -restricted smoothness (RSM) and m -restricted strong convexity (RSC), the set function $f(\mathbf{S}) = -l(\boldsymbol{\beta}_{\mathbf{S}})$ is weakly submodular with parameter $\gamma \geq m/M$. The parameters M , m , and γ are defined formally in Section 2. We use this result to analyze three greedy algorithms, each progressively better but more computationally intensive: the Oblivious (or Marginal Regression) algorithm computes for each feature the increase in objective and keeps the k individually best features. Orthogonal Matching Pursuit (OMP) greedily adds one feature at a time by picking the feature with the largest inner product with the function gradient. The gradient is the correct generalization of the residual error used in linear regression OMP. Finally, the most sophisticated algorithm is Forward Stepwise Regression: it adds one feature at a time by re-fitting the model repeatedly and keeping the feature that best improves the objective function at each step. We obtain the following performance bounds:

- The Oblivious algorithm produces a (γ/k) -approximation to the best k -subset after k steps.
- Orthogonal Matching Pursuit produces a $(1 - e^{-m/M})$ -approximation to the best k -subset after k steps.
- Forward Stepwise Regression produces a $(1 - e^{-\gamma})$ -approximation to the best k -subset after k steps.

We also show that if Forward Stepwise Regression is used to select more than k features, we can approximate the best k -sparse feature performance within an *arbitrary* accuracy. Finally, under additional assumptions, we derive statistical guarantees for convergence of the greedily selected parameter to the optimal sparse solution. Note that our results yield stronger performance guarantees even for cases that have been previously studied under the same assumptions. For example, for linear regression we obtain a better exponent in the approximation factor of OMP compared to previous state-of-the-art [13] (see Remark 8).

One implication of our work is that weak submodularity seems to be a sharper technical tool than RSC, as any function satisfying the latter also satisfies the former. Das and Kempe [13] noted that it is easy to find problems which satisfy weak submodularity but not RSC, emphasizing the limitations of spectral techniques versus submodularity. We show this holds beyond linear regression, for any likelihood function.

Our connection between restricted strong convexity and weak submodularity has many benefits. First, the weak submodularity framework can now be used to develop theory for additional problems where spectral conditions would be overly restrictive or unwarranted. In our ongoing work, for example, we show that RSC assumptions play an important role in characterizing distributed greedy maximization [27] and rank constrained matrix optimization [26]. Additionally, [16] studies the role of RSC in streaming feature selection, as well as interpretability of black-box neural network models. Second, this framework allows statisticians to draw from classical results and recent advances in the field of (weak) submodular set

function theory to provide general guarantees on the performance of greedy algorithms [6, 11].

Related work. There have been a wide range of techniques developed for solving problems with sparsity constraints. These include using the Lasso, greedy selection methods (such as Forward Stagewise/Stepwise Regressions [35], OMP, and CoSaMP [38]), forward–backward methods [24, 31], Pareto optimization [43], exponentially weighted aggregation [45], and truncated gradient methods [23]. Under the restricted strong convexity and smoothness assumptions that will be outlined in the next section, forward–backward methods can in fact recover the correct support of the optimal set of parameters under an assumption on the smallest value of the optimal variable as it relates to the gradient. In contrast, the results derived in our setting for sparse GLMs allow one to provide recovery guarantees at various sparsity levels regardless of the optimal solution, with only information on the desired sparsity level and the RSC and RSM parameters. This is again in contrast to the other work that also needs information on the smallest nonzero value in the optimal set of coefficients, as well as an upper bound on the gradient of the objective at this optimal set.

Focusing explicitly on OMP, most previous results require the strong RIP assumption, whereas we only require the weaker RSC and RSM assumptions. In our setting of arbitrary model conditions, OMP requires RIP as highlighted in Corollary 2.1 of Zhang [53]. However, we do note that under certain stochastic assumptions, for instance independent noise, the results established in those works can provide sharper guarantees with respect to the number of samples required by a factor on the order of $\log[(k \log p)/n]$ (see Section 4). Nevertheless, we emphasize that our results apply under arbitrary assumptions on the noise and use only RSC and RSM assumptions.

In [13], Das and Kempe’s framework optimizes the goodness of fit parameter R^2 in linear regression. We derive similar results without relying on the closed-form solution to least squares. Greedy algorithms are prevalent in compressed sensing literature [38] and statistical learning theory [5]. Greedy methods for sparsity constrained regression were analyzed in [23, 24, 31, 34, 46, 51] under assumptions similar to ours but without connections to submodularity. Convergence guarantees for ℓ_1 regularized regression were given for exponential families in [25], and for general nonlinear functions in [50]. However, the latter requires additional assumptions such as knowledge of the nonlinearity and bounds on the loss function’s derivatives, which can again be derived under appropriate stochasticity and model assumptions.

Classical results on submodular optimization [12, 41] typically do not scale to large-scale applications. Therefore, several recent algorithms improve efficiency at the expense of slightly weaker guarantees [4, 16, 27, 36, 42]. Submodularity has been used recently in the context of active learning [20, 49]. In this setup, the task is to select predictive data points instead of features. Recently, [1] and [21]

obtained constant factor guarantees for greedy algorithms using techniques from submodularity even though the problems considered were not strictly submodular. In [6], the authors show tight bounds for maximizing cardinality constrained, weakly submodular set functions that also have bounded curvature. For maximizing weakly submodular functions subject to more general matroid constraints, [11] recently proved that a randomized greedy forward stepwise algorithm has a constant factor approximation guarantee.

There are deep connections between convexity and submodularity [2]. For example, the convex closure of a submodular function can be tractably computed as its Lovász extension [33]. This connection is fundamental to providing polynomial-time minimization algorithms for submodular set functions [17, 19]. Similarly, another continuous extension of set functions, called the multilinear extension, is vital for algorithmic development of constant factor approximation guarantees for submodular maximization [10]. A more detailed study of convexity/concavity-like properties of submodular functions was presented in [22]. More recent works exploit similar connections to provide constant factor approximation guarantees for a class of nonconvex functions [7, 15, 18].

2. Preliminaries. First, we collect some notation that will be used throughout the remainder of this paper. Sets are represented by sans script fonts, for example, A, B . Vectors are represented using lower case bold letters for example, \mathbf{x}, \mathbf{y} and matrices are represented using upper case bold letters, for example, \mathbf{X}, \mathbf{Y} . The i th column of \mathbf{X} is denoted \mathbf{X}_i . Nonbold face letters are used for scalars, for example, j, M, r and function names, for example, $f(\cdot)$. The transpose of a vector or a matrix is represented by \top , for example, \mathbf{X}^\top . For any vector \mathbf{v} , define $\|\mathbf{v}\|_{2,k} = \sqrt{\sum_{i=1}^k v_{(i)}^2}$, where $v_{(i)}$ represent the order statistics of \mathbf{v} in decreasing order. Define $[p] := \{1, 2, \dots, p\}$. For simplicity, we assume a set function defined on a ground set of size p has domain $2^{[p]}$. For singleton sets, we write $f(j) := f(\{j\})$.

Recall that a set function $f(\cdot) : 2^{[p]} \rightarrow \mathbb{R}$ is called submodular if and only if for all $A, B \subseteq [p]$, $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$. Intuitively, submodular functions have a *diminishing returns* property. This becomes clear when A and B are disjoint: the sets have less value taken together than they have individually. We also state an equivalent definition.

DEFINITION 1 (Proposition 2.3 in [2]). $f(\cdot)$ is submodular if for all $A \subseteq [p]$ and $j, k \in [p] \setminus A$,

$$f(A \cup \{k\}) - f(A) \geq f(A \cup \{j, k\}) - f(A \cup \{j\}).$$

The function is called normalized if $f(\emptyset) = 0$ and monotone if and only if $f(A) \leq f(B)$ for all $A \subseteq B$. A seminal result by Nemhauser et al. [41] shows that greedy maximization of a monotone, submodular function (Algorithm 2 in Section 2.2) returns a set with value within a factor of $(1 - 1/e)$ from the optimum set

of the same size. This has been the starting point for several algorithmic advances for large-scale combinatorial optimization, including stochastic, distributed, and streaming algorithms [4, 16, 27, 36].

Next, we define the submodularity ratio of a monotone set function.

DEFINITION 2 (Submodularity ratio [13], weak submodularity). Let $S, L \subset [p]$ be two disjoint sets, and $f(\cdot) : 2^{[p]} \rightarrow \mathbb{R}$. The submodularity ratio of L with respect to S is given by

$$(2) \quad \gamma_{L,S} := \frac{\sum_{j \in S} [f(L \cup \{j\}) - f(L)]}{f(L \cup S) - f(L)}.$$

The submodularity ratio of a set U with respect to an integer k is given by

$$(3) \quad \gamma_{U,k} := \min_{\substack{L,S:L \cap S = \emptyset, \\ L \subseteq U, |S| \leq k}} \gamma_{L,S}.$$

Let $\gamma > 0$. We call a function γ -weakly submodular at a set U and an integer k if $\gamma_{U,k} \geq \gamma$.

It is straightforward to show that $f(\cdot)$ is submodular if and only if $\gamma_{L,S} \geq 1$ for all sets L and S . In our application, $0 < \gamma_{L,S} \leq 1$ which provides a notion of *weak submodularity* in the sense that even though the function is not submodular, it still provides provable bounds of performance of greedy selections.

Next, we define the restricted versions of strong concavity and smoothness, consistent with [32, 40].

DEFINITION 3 (Restricted strong concavity, restricted smoothness). A function $l : \mathbb{R}^p \rightarrow \mathbb{R}$ is said to be restricted strong concave with parameter m_Ω and restricted smooth with parameter M_Ω on a domain $\Omega \subset \mathbb{R}^p \times \mathbb{R}^p$ if for all $(\mathbf{x}, \mathbf{y}) \in \Omega$,

$$-\frac{m_\Omega}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \geq l(\mathbf{y}) - l(\mathbf{x}) - \langle \nabla l(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq -\frac{M_\Omega}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

REMARK 1. If a function $l(\cdot)$ has restricted strong concavity parameter m , then its negative $-l(\cdot)$ has restricted strong convexity parameter m . In the sequel, we will use these properties interchangeably for maximum likelihood estimation where $l(\cdot)$ is the log-likelihood function and $-l(\cdot)$ is the data fit loss.

If $\Omega' \subseteq \Omega$, then $M_{\Omega'} \leq M_\Omega$ and $m_{\Omega'} \geq m_\Omega$. With slight abuse of notation, unless stated otherwise let (m_k, M_k) denote the RSC and RSM parameters on the domain Ω_k of all pairs of k -sparse vectors that differ in at most k entries, that is, $\Omega_k := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x}\|_0 \leq k, \|\mathbf{y}\|_0 \leq k, \|\mathbf{x} - \mathbf{y}\|_0 \leq k\}$. If $j \leq k$, then $M_j \leq M_k$ and $m_j \geq m_k$. In addition, denote $\tilde{\Omega}_k := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x}\|_0 \leq k, \|\mathbf{y}\|_0 \leq k, \|\mathbf{x} - \mathbf{y}\|_0 \leq 1\}$ with corresponding smoothness parameter \tilde{M}_k , which is clearly greater than or equal to $\tilde{M}_1 = M_1$.

2.1. *Sparsity constrained generalized linear regression.* Due to its combinatorial nature, there has been a tremendous amount of effort in developing computationally tractable and fundamentally sound methods to solve the subset selection problem approximately. In this section, we provide background on various problems that arise in subset selection. Our focus here will be on sparse regression problems. We will assume that we obtain n observations of the form (\mathbf{x}_i, y_i) . For now, we make no assumptions regarding how the data is generated, but wish to model the interaction between $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ as

$$y_i = g(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle) + \text{noise},$$

for some known link function g and a sparse vector $\boldsymbol{\beta}^*$. Each feature observation is a row in the $n \times p$ design matrix \mathbf{X} . The above is called a generalized linear model, or GLM, and arises as the maximum likelihood estimate of data drawn from a canonical exponential family, that is, normal, Bernoulli, Dirichlet, negative binomial, etc. [9]. Another interpretation is in minimizing the average Bregman divergence between the response y_i and the mean parameter $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$. There has been a large body of literature studying this method’s statistical properties. These include establishing sparsistency, parameter consistency, and prediction error [25, 40, 47]. We refer the reader to the standard literature for more details on GLMs and exponential families [9, 14].

2.2. *Support selection algorithms.* We study general M -estimators of the form (1) for some function $l(\cdot)$. Note that $l(\cdot)$ will implicitly depend on our specific data set, but we hide that for ease of notation. One common choice of $l(\cdot)$ is the log-likelihood of a parametric distribution. Das and Kempe [13] consider the specific case of maximizing the R^2 objective [13]. Through a simple transformation, that is equivalent to maximizing the log-likelihood of the parametric distribution that arises from the model $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle + w$ where $w \sim N(0, \sigma^2)$. If we let $\widehat{\boldsymbol{\beta}}^s$ be the s -sparse solution derived, and again let $\bar{\boldsymbol{\beta}}^k$ be the best k -sparse parameter, then we wish to bound

$$l(\widehat{\boldsymbol{\beta}}^s) \geq (1 - \varepsilon)l(\bar{\boldsymbol{\beta}}^k),$$

without any assumptions on the underlying sparsity or a *true* parameter.

For a concave function $l(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$, we can define an equivalent set function $\bar{f}(\cdot) : 2^{[p]} \rightarrow \mathbb{R}$ so that $\bar{f}(\mathbf{S}) = \max_{\text{supp}(\mathbf{x}) \subseteq \mathbf{S}} l(\mathbf{x})$. The problem of support selection for a given integer k is then $\max_{|\mathbf{S}| \leq k} \bar{f}(\mathbf{S})$. Recall that a vector is k -sparse if it is 0 on all but k indices. We provide approximation guarantees on the *normalized* set function defined as $f(\mathbf{S}) = \bar{f}(\mathbf{S}) - \bar{f}(\emptyset)$. The support selection problem is thus equivalent to finding the k -sparse vector $\boldsymbol{\beta}$ that maximizes $l(\boldsymbol{\beta})$:

$$(4) \quad \max_{\mathbf{S}:|\mathbf{S}| \leq k} f(\mathbf{S}) \Leftrightarrow \max_{\substack{\boldsymbol{\beta}: \boldsymbol{\beta}^{\mathbf{S}^c} = 0 \\ |\mathbf{S}| \leq k}} l(\boldsymbol{\beta}) - l(\mathbf{0}).$$

Algorithm 1 Oblivious support selection

- 1: **Input:** sparsity parameter k , set function $f(\cdot) : 2^{[p]} \rightarrow \mathbb{R}$
- 2: **for** $i = 1 \dots p$ **do**
- 3: $\mathbf{v}[i] \leftarrow f(\{i\})$
- 4: **end for**
- 5: $\mathbf{S}_k \leftarrow$ indices corresponding to the top k values of \mathbf{v}
- 6: **return** $\mathbf{S}_k, f(\mathbf{S}_k)$.

Let $\boldsymbol{\beta}^{(A)}$ denote the $\boldsymbol{\beta}$ maximizing $f(A)$, and let $\boldsymbol{\beta}_B^{(A)}$ denote $\boldsymbol{\beta}^{(A)}$ restricted to the coordinates specified by B . We present three support selection strategies for the set function $f(\cdot)$ that are simple to implement and are widely used.

Oblivious algorithm. One natural strategy is to select the top k features ranked by their individual improvement over a null model, using a goodness of fit metric such as R^2 or p -value. This is referred to as the *Oblivious* algorithm, shown as Algorithm 1. In the context of linear regression, this is simply Marginal Regression. While it is computationally inexpensive and parallelizes easily, the Oblivious algorithm does not account for dependencies or redundancies in the span of features.

Forward stepwise algorithm. A less extreme greedy approach would check for incremental gain at each step using nested models. This is referred to as the *Forward Stepwise* algorithm, presented as Algorithm 2. Given a set of features \mathbf{S} is already selected, choose the feature with largest marginal gain, that is, select $\{j\}$ such that $\mathbf{S} \cup \{j\}$ has the most improvement over \mathbf{S} . All regression coefficients are updated each time a new feature is added. In the case of submodular set functions, this returns a solution that is provably within a constant factor of the optimum [41].

Generalized OMP. Another approach is to choose features which correlate well with the orthogonal complement of what has already been selected. Using (4) and an appropriately chosen model, we can define the gradient evaluated at the current parameter $\boldsymbol{\beta}$ to be a residual term. In *Orthogonal Matching Pursuit*,

Algorithm 2 Forward stepwise selection

- 1: **Input:** sparsity parameter k , set function $f(\cdot) : 2^{[p]} \rightarrow \mathbb{R}$
- 2: $\mathbf{S}_0^G \leftarrow \emptyset$
- 3: **for** $i = 1 \dots k$ **do**
- 4: $s \leftarrow \arg \max_{j \in [p] \setminus \mathbf{S}_{i-1}} f(\mathbf{S}_{i-1}^G \cup \{j\}) - f(\mathbf{S}_{i-1}^G)$
- 5: $\mathbf{S}_i^G \leftarrow \mathbf{S}_{i-1}^G \cup \{s\}$
- 6: **end for**
- 7: **return** $\mathbf{S}_k^G, f(\mathbf{S}_k^G)$.

Algorithm 3 Orthogonal matching pursuit

```

1: Input: sparsity parameter  $k$ , objective function  $l(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$ 
2:  $\mathbf{S}_0^P \leftarrow \emptyset$ 
3:  $\mathbf{r} \leftarrow \nabla l(\mathbf{0})$ 
4: for  $i = 1 \dots k$  do
5:    $s \leftarrow \arg \max_j |\langle \mathbf{e}_j, \mathbf{r} \rangle|$ 
6:    $\mathbf{S}_i^P \leftarrow \mathbf{S}_{i-1}^P \cup \{s\}$ 
7:    $\boldsymbol{\beta}^{(\mathbf{S}_i^P)} \leftarrow \arg \max_{\boldsymbol{\beta} : \text{supp}(\boldsymbol{\beta}) \subseteq \mathbf{S}_i^P} l(\boldsymbol{\beta})$ 
8:    $\mathbf{r} \leftarrow \nabla l(\boldsymbol{\beta}^{(\mathbf{S}_i^P)})$ 
9: end for
10: return  $\mathbf{S}_k^P, l(\boldsymbol{\beta}^{(\mathbf{S}_k^P)})$ 

```

features are selected to maximize the inner product with this residual, as shown in line 5 of Algorithm 3. Here, \mathbf{e}_j represents a unit vector with a 1 in coordinate j and zeros in the other $p - 1$ coordinates. OMP requires much less computation than forward stepwise selection, since the feature comparison is done via an n -dimensional inner product rather than a regression score. A detailed discussion can be found in [8].

3. Approximation guarantees. In this section, we derive theoretical lower bounds on the submodularity ratio based on strong concavity and strong smoothness of a function $l(\cdot)$. We show that if the concavity parameter is bounded away from 0 and the smoothness parameter is finite, then the submodularity ratio is also bounded away from 0, which allows approximation guarantees for Algorithms 1–3. While our proof techniques differ substantially, the outline of this section follows that of [13] which obtained approximation guarantees for support selections for linear regression. While our results are applicable to general functions, in Appendix B we discuss a direct application of maximum likelihood estimation for sparse generalized linear models.

We assume a differentiable function $l : \mathbb{R}^p \rightarrow \mathbb{R}$. Recall that we can define the equivalent, normalized, monotone set function $f : 2^{[p]} \rightarrow \mathbb{R}$ for a selected support as $f(\mathbf{S}) = \max_{\text{supp}(\boldsymbol{\beta}) \subseteq \mathbf{S}} l(\boldsymbol{\beta}) - l(\mathbf{0})$. We will use set functions wherever possible to simplify the notation.

We now present our main result as Theorem 1, a bound on a function's submodularity ratio $\gamma_{\mathcal{U},k}$ in terms of its strong concavity and smoothness parameters (see Definitions 2–3). Proofs of lemmas and theorems omitted from this section can be found in Section 5.

THEOREM 1 (RSC/RSM implies weak submodularity). *Define $f(\mathbf{S})$ as in (4), with a function $l(\cdot)$ that is $(m_{|\mathcal{U}|+k}, M_{|\mathcal{U}|+k})$ -(strongly concave, smooth) on $\Omega_{|\mathcal{U}|+k}$*

and $\tilde{M}_{|\mathbf{U}|+1}$ smooth on $\tilde{\Omega}_{|\mathbf{U}|+1}$. Then the submodularity ratio $\gamma_{\mathbf{U},k}$ is lower bounded by

$$(5) \quad \gamma_{\mathbf{U},k} \geq \frac{m_{|\mathbf{U}|+k}}{\tilde{M}_{|\mathbf{U}|+1}} \geq \frac{m_{|\mathbf{U}|+k}}{M_{|\mathbf{U}|+k}}.$$

REMARK 2. In the case of linear least-squares regression, m and M become sparse eigenvalues of the covariance matrix, that is, $m_{|\mathbf{U}|+k} = \lambda_{\min}(|\mathbf{U}| + k) \geq 0$ and $\tilde{M}_{|\mathbf{U}|+1} = \lambda_{\max}(1) = 1$. Thus Theorem 1 becomes $\gamma_{\mathbf{U},k} \geq \lambda_{\min}(|\mathbf{U}| + k)$, that is, “RIP implies weak submodularity,” consistent with Lemma 2.4 of [13].

REMARK 3. Since $m/M \leq 1$, this method cannot prove that the function is submodular (even on a restricted set of features). However, the guarantees in this section only require weak submodularity.

REMARK 4. Theorem 1 has the following geometric interpretation: the submodularity ratio of $f(\mathbf{S})$ is bounded in terms of the maximum curvature of $l(\cdot)$ over the domain $\tilde{\Omega}_{|\mathbf{U}|+1}$ and the minimum curvature of $l(\cdot)$ over the (larger) domain $\Omega_{|\mathbf{U}|+k}$. The upper-curvature bound effectively controls the maximum amount that each individual function coordinate can influence the function value. The lower-curvature bound provides a lower-bound on the improvement of adding all features at once. Hence, loosely speaking the submodular ratio bound will be the ratio of both of these quantities. This intuition is used more formally in the proof (Section 5), where for a k -sparse set \mathbf{S} and $j \in \mathbf{S}$, $l(\boldsymbol{\beta}^{(\mathbf{U})})$ is perturbed by scaled projections of $\nabla l(\boldsymbol{\beta}^{(\mathbf{U})})$ onto $\boldsymbol{\beta}_j^{(\mathbf{U},\mathbf{S})}$ and $\mathbf{e}_{\mathbf{S}}$, respectively.

Theorem 1 allows us to generalize several results of [13], starting with the following lemma.

LEMMA 1. Let $1 \leq k \leq n$:

$$\begin{aligned} f([k]) &\geq \max \left\{ \frac{1}{k}, \frac{m_1}{4M_k} \left(3 + \frac{m_1}{M_1} \right) \right\} \sum_{j=1}^k f(j) \\ &\geq \max \left\{ \frac{1}{k}, \frac{m_k}{4M_k} \left(3 + \frac{m_k}{M_k} \right) \right\} \sum_{j=1}^k f(j). \end{aligned}$$

Now we present our first performance guarantee for feature selection.

THEOREM 2 (Oblivious algorithm guarantee). Define $f(\mathbf{S})$ as in (4), with a function $l(\cdot)$ that is M_k -smooth and m_k -strongly concave on Ω_k . Let f^{OBL} be the

value at the set selected by the Oblivious algorithm, and let f^{OPT} be the optimal value over all sets of size k . Then

$$\begin{aligned} f^{\text{OBL}} &\geq \max \left\{ \frac{m_k}{kM_1}, \frac{m_k m_1}{4M_k M_1} \left(3 + \frac{m_1}{M_1} \right) \right\} f^{\text{OPT}} \\ &\geq \max \left\{ \frac{m_k}{kM_k}, \frac{3m_k^2}{4M_k^2}, \frac{m_k^3}{M_k^3} \right\} f^{\text{OPT}}. \end{aligned}$$

REMARK 5. When the function is modular, that is, $m_\Omega = M_\Omega$ for all Ω , then $f^{\text{OBL}} = f^{\text{OPT}}$ and the bound in Theorem 2 holds with equality.

Next, we prove a stronger, constant factor approximation guarantee for the greedy, Forward Stepwise algorithm.

THEOREM 3 (Forward stepwise algorithm guarantee). Define $f(\mathbf{S})$ as in (4), with a function that is M -smooth and m -strongly concave on Ω_{2k} . Let \mathbf{S}_k^G be the set selected by the FS algorithm and \mathbf{S}^* be the optimal set of size k corresponding to values f^{FS} and f^{OPT} . Then

$$(6) \quad f^{\text{FS}} \geq (1 - e^{-\gamma_{\mathbf{S}_k^G, k}}) f^{\text{OPT}} \geq (1 - e^{-m/M}) f^{\text{OPT}}.$$

REMARK 6. This constant factor bound can be improved by running the Forward Stepwise algorithm for $r > k$ steps. The proof of Theorem 3 generalizes to compare performance of r greedy iterations to the optimal k -subset of features. This generalized bound does not necessarily approach 1 as $r \rightarrow \infty$, however, since $\gamma_{\mathbf{S}_r^G, k}$ is a decreasing function of r .

COROLLARY 1. Let $f^{\text{FS}+}$ denote the solution obtained after r iterations of the Forward Stepwise algorithm, and let f^{OPT} be the objective at the optimal k -subset of features. Let $\gamma = \gamma_{\mathbf{S}_r^G, k}$ be the submodularity ratio associated with the output of $f^{\text{FS}+}$ and k . Then

$$f^{\text{FS}+} \geq (1 - e^{-\gamma(r/k)}) f^{\text{OPT}}.$$

In particular, setting $r = ck$ corresponds to a $(1 - e^{-c\gamma})$ -approximation, and setting $r = k \log n$ corresponds to a $(1 - n^{-\gamma})$ -approximation.

Corollary 1 is useful when γ can be bounded on larger support sets. We next present approximation guarantees when γ can only be bounded on smaller support sets.

THEOREM 4. Define $f(\mathbf{S})$ as in (4), with a function $l(\cdot)$ that is m' -strongly concave on Ω_k and M' -smooth on $\tilde{\Omega}_k$. Let \mathbf{S}_k^G be the set of features selected by

the Forward Stepwise algorithm and S_k be the optimal feature set on k variables corresponding to values f^G and f^{OPT} . Then

$$f^{\text{FS}} \geq \Theta(2^{-M'/m'})(1 - e^{-m'/M'}) f^{\text{OPT}}.$$

REMARK 7. We note that our bounds are loose for certain special cases like modular functions and linear regression. These require making use of additional tools and specific properties of the function and data at hand (see [13]).

Orthogonal Matching Pursuit is more computationally efficient than forward stepwise regression, since step i only fits one regression instead of $p - i$. Thus we have a weaker guarantee than Theorem 3. Similar to Corollary 1, this result generalizes to running OMP for $r > k$ iterations.

THEOREM 5 (OMP algorithm guarantee). Define $f(S)$ as in (4), with a log-likelihood function that is (M, m) -(smooth, strongly concave) on Ω_{2k} . Let f^{OMP} be the value at the set of features selected by the OMP algorithm and f^{OPT} be the optimal value over all sets of size k . Then

$$f^{\text{OMP}} \geq (1 - e^{-m/M}) f^{\text{OPT}}.$$

COROLLARY 2. Let f^{P+} denote the solution obtained after r iterations of the OMP algorithm, and let f^{OPT} be the objective at the optimal k -subset of features. Let $\alpha = (m/M)$ be the ratio associated with the output of f^{P+} and k . Then

$$f^{P+} \geq (1 - e^{-\alpha(r/k)}) f^{\text{OPT}}.$$

In particular, setting $r = ck$ corresponds to a $(1 - e^{-c\alpha})$ -approximation, and setting $r = k \log n$ corresponds to a $(1 - n^{-\alpha})$ -approximation.

REMARK 8. Theorem 5 and Corollary 2 improve on the approximation guarantee of [13] by a factor of γ in the exponent. Previous work obtained the approximation factor $1 - e^{-\gamma \lambda_{\min}(2k)}$, whereas the proof of Theorem 5 establishes $1 - e^{-\lambda_{\min}(2k)}$. Therefore, we obtain a better exponent for linear regression and also generalize to any likelihood function. Theorem 3 also gives intuition on when the performance of OMP will differ from that of Forward Selection, that is, when the inequality (6) is loose.

4. Statistical recovery guarantees. Understanding optimization guarantees are useful, but do not clearly translate to bounds on parameter recovery. Below we present a general theorem that allows us to derive parameter bounds. When combined with Section 3, it produces recovery guarantees for greedy algorithms as special cases.

THEOREM 6 (Parameter recovery guarantees). *Suppose that after r iterations to approximate a function evaluated at a set \mathbf{S}_s^* of cardinality s , we have the guarantee that*

$$f(\mathbf{S}_r) \geq C_{s,r} f(\mathbf{S}_s^*).$$

Recall that $f(\mathbf{S}_r) = \max_{\text{supp}(\boldsymbol{\beta}) \subset \mathbf{S}_r} l(\boldsymbol{\beta}) - l(\mathbf{0})$. Let $\widehat{\boldsymbol{\beta}}^r$ be the solution to the optimization problem and consider any arbitrary s -sparse vector $\boldsymbol{\beta}^s$ with support on \mathbf{S}_s^ . Then, under m_{s+r} RSC on Ω_{s+r} we have that*

$$\|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^s\|_2^2 \leq \frac{4}{m_{s+r}^2} \|\nabla l(\boldsymbol{\beta}^s)\|_{2,(s+r)}^2 + \frac{4}{m_{s+r}} (1 - C_{s,r}) [l(\boldsymbol{\beta}^s) - l(\mathbf{0})].$$

For the remainder of this section, we consider several cases of Theorem 6 and compare to results from previous work.

4.1. Forward selection with linear regression model. First, we will consider a special case of Algorithm 2 for linear regression where the rows of the design matrices are $N(0, \Sigma)$ for a covariance matrix of the form $\Sigma = \mathbf{I} + \mathbf{1}\mathbf{1}^T$. Further, we assume the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w},$$

where $\|\boldsymbol{\beta}^*\|_2 \leq 1$ and is s -sparse, the rows of $\mathbf{X} \in \mathbb{R}^{n \times p}$ are $N(0, \Sigma)$, and $w_i \sim N(0, \sigma^2)$ are i.i.d. We also take $l(\boldsymbol{\beta}) = \frac{1}{n} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

COROLLARY 3. *Given the above setup, if $(s + r)\sigma^2 \log p = o(n)$ and $r = \Omega(s \log n)$, then the parameter error goes to zero with high probability as $n \rightarrow \infty$.*

4.2. Orthogonal matching pursuit with linear regression model. Next, we consider the results of Zhang, which provides parameter recovery bounds in the case of OMP (Algorithm 3). The simplest comparison is to contrast our results with Corollary 2.2 of [53]. Consider the linear regression model above with an original s -sparse vector, r iterations of the algorithm, and a spiked identity covariance model, $\Sigma = (1 - a)\mathbf{I} + a\mathbf{1}\mathbf{1}^T$.

PROPOSITION 1. *While Theorem 6 holds for any a , Zhang [53] requires that a does not exceed $\frac{1}{s+1}$.*

PROOF. Zhang requires the RIP condition to hold, namely $M_s \leq 2m_{s+r}$. We know that the difference between means of $2\lambda_{\min}(s + r)$ and $\lambda_{\max}(s)$ is $\Delta = 1 - a - as$. Since $\Delta/2 \leq \mu/2$ in both cases and χ^2 variables concentrate within constant factors of their means, we have $M_s \leq \frac{3}{2}(1 - a) + \frac{as}{2} \leq 2m_{s+r}$. However, $\Delta > 0 \Leftrightarrow s \leq \frac{1}{a} - 1$. Rearranging, we have $a \leq \frac{1}{s+1}$. Thus, as has been

noted in prior work, the RIP condition will not hold for the spiked model in settings where a is much larger than $\frac{1}{s+1}$. \square

Nevertheless, we can still proceed and assume that the RIP condition is not required. In that case, the bound established in [53] shows

$$\|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2^2 \leq 24M_{s+r} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\|_2^2 / m_{s+r}^2.$$

When M_{s+r} and m_{s+r} are of the same order, then this result is better than ours by log factors. However, when we consider a case like the spiked covariance model, then our results are better by a factor of s in terms of statistical accuracy, but worse by a factor of $\log n$ with respect to sample complexity.

4.3. *Orthogonal matching pursuit with logistic regression model.* Finally, consider our bounds for OMP (Algorithm 3) in the case of logistic regression. Applying our approximation guarantees in Theorem 5 matches the bound given by Theorem 2 of [34] up to constant factors. However, their guarantee for parameter recovery requires a condition that is only known to be satisfied under incoherence assumptions. Our Theorem 6 holds more generally. Their conditions on exact recovery are incomparable with our statistical error bounds.

5. Theorem proofs.

5.1. *Proof of Theorem 1.*

PROOF. We proceed by upper bounding the denominator and lower bounding the numerator of (2). Let $\bar{k} = |\mathbf{L}| + k$. First, we apply Definition 3 with $\mathbf{x} = \boldsymbol{\beta}^{(\mathbf{L})}$ and $\mathbf{y} = \boldsymbol{\beta}^{(\mathbf{LUS})}$,

$$(7) \quad \frac{m_{\bar{k}}}{2} \|\boldsymbol{\beta}^{(\mathbf{LUS})} - \boldsymbol{\beta}^{(\mathbf{L})}\|_2^2 \leq l(\boldsymbol{\beta}^{(\mathbf{L})}) - l(\boldsymbol{\beta}^{(\mathbf{LUS})}) + \langle \nabla l(\boldsymbol{\beta}^{(\mathbf{L})}), \boldsymbol{\beta}^{(\mathbf{LUS})} - \boldsymbol{\beta}^{(\mathbf{L})} \rangle.$$

Rearranging and noting that $l(\cdot)$ is monotone for increasing supports,

$$\begin{aligned} 0 &\leq l(\boldsymbol{\beta}^{(\mathbf{LUS})}) - l(\boldsymbol{\beta}^{(\mathbf{L})}) \leq \langle \nabla l(\boldsymbol{\beta}^{(\mathbf{L})}), \boldsymbol{\beta}^{(\mathbf{LUS})} - \boldsymbol{\beta}^{(\mathbf{L})} \rangle - \frac{m_{\bar{k}}}{2} \|\boldsymbol{\beta}^{(\mathbf{LUS})} - \boldsymbol{\beta}^{(\mathbf{L})}\|_2^2 \\ &\leq \max_{\mathbf{v}: \mathbf{v}_{(\mathbf{LUS})^c} = 0} \langle \nabla l(\boldsymbol{\beta}^{(\mathbf{L})}), \mathbf{v} - \boldsymbol{\beta}^{(\mathbf{L})} \rangle - \frac{m_{\bar{k}}}{2} \|\mathbf{v} - \boldsymbol{\beta}^{(\mathbf{L})}\|_2^2. \end{aligned}$$

Setting $\mathbf{v} = \boldsymbol{\beta}^{(\mathbf{L})} + \frac{1}{m_{\bar{k}}} \nabla l(\boldsymbol{\beta}^{(\mathbf{L})})_{\mathbf{S}}$, we have

$$(8) \quad 0 \leq l(\boldsymbol{\beta}^{(\mathbf{LUS})}) - l(\boldsymbol{\beta}^{(\mathbf{L})}) \leq \frac{1}{2m_{\bar{k}}} \|\nabla l(\boldsymbol{\beta}^{(\mathbf{L})})_{\mathbf{S}}\|_2^2.$$

Next, consider a single coordinate $j \in \mathbf{S}$. The function at $\boldsymbol{\beta}^{(\mathbf{LU}\{j\})}$ is larger than the function at any other $\boldsymbol{\beta}$ on the same support. In particular $l(\boldsymbol{\beta}^{(\mathbf{LU}\{j\})}) \geq l(\mathbf{y}_j)$,

where $\mathbf{y}_j := \boldsymbol{\beta}^{(L)} + \alpha_j \boldsymbol{\beta}_j^{(LUS)}$ for some scalar α_j . Noting that $(\mathbf{x} = \boldsymbol{\beta}^{(L)}, \mathbf{y} = \mathbf{y}_j) \in \tilde{\Omega}_{|L|+1}$ and applying Definition 3,

$$\begin{aligned} l(\boldsymbol{\beta}^{(LU\{j\})}) - l(\boldsymbol{\beta}^{(L)}) &\geq l(\boldsymbol{\beta}^{(L)} + \alpha_j \boldsymbol{\beta}_j^{(LUS)}) - l(\boldsymbol{\beta}^{(L)}) \\ &\geq \langle \nabla l(\boldsymbol{\beta}^{(L)}), \alpha_j \boldsymbol{\beta}_j^{(LUS)} \rangle - \frac{\tilde{M}_{|L|+1}}{2} |\alpha_j \boldsymbol{\beta}_j^{(LUS)}|^2. \end{aligned}$$

Summing over all $j \in S$ and setting

$$\alpha_j = \frac{\langle \nabla l(\boldsymbol{\beta}^{(L)}), \boldsymbol{\beta}_j^{(LUS)} \rangle}{\tilde{M}_{|L|+1} |\boldsymbol{\beta}_j^{(LUS)}|^2},$$

we have

$$\begin{aligned} l(\boldsymbol{\beta}^{(LU\{j\})}) - l(\boldsymbol{\beta}^{(L)}) &\geq \frac{(\langle \nabla l(\boldsymbol{\beta}^{(L)}), \boldsymbol{\beta}_j^{(LUS)} \rangle)^2}{2\tilde{M}_{|L|+1} |\boldsymbol{\beta}_j^{(LUS)}|^2} \\ \Rightarrow \sum_{j \in S} l(\boldsymbol{\beta}^{(LU\{j\})}) - l(\boldsymbol{\beta}^{(L)}) &\geq \frac{1}{2\tilde{M}_{|L|+1}} \sum_{j \in S} (\nabla l(\boldsymbol{\beta}^{(L)}))_j^2 \\ &= \frac{1}{2\tilde{M}_{|L|+1}} \|\nabla l(\boldsymbol{\beta}^{(L)})_S\|_2^2. \end{aligned}$$

Substituting the above line and (8) into (2), the result follows from taking the minimum over all sets L, S . \square

5.2. Proof of Lemma 1.

PROOF. Let $S = [k]$. Since $f(\cdot)$ is monotone, $f(j) \leq f(S)$ for $j \in S$. Summing over all $j \in S$ and dividing by k yields the first part of the inequality. The rest of the proof requires combining several applications of Definition 3 to the underlying likelihood function l for carefully chosen \mathbf{x}, \mathbf{y} . Define a k -sparse $\bar{\boldsymbol{\beta}}$ by $\bar{\boldsymbol{\beta}}_j = \alpha_j \boldsymbol{\beta}_j^{(j)}, j \in S$ for some positive scalar α_j and 0 elsewhere. First, we apply Definition 3 with $\mathbf{x} = \mathbf{0}$ and $\mathbf{y} = \bar{\boldsymbol{\beta}}$. This implies

$$(9) \quad l(\bar{\boldsymbol{\beta}}) \geq \langle \nabla l(\mathbf{0}), \bar{\boldsymbol{\beta}} \rangle - \frac{M_k}{2} \sum_{j \in S} |\alpha_j \boldsymbol{\beta}_j^{(j)}|^2.$$

Next, applying the same definition k times with $\mathbf{x} = \mathbf{0}$ and $\mathbf{y} = \boldsymbol{\beta}^{(j)}$ and summing over $j \in S$,

$$\begin{aligned} (10) \quad \langle \nabla l(\mathbf{0}), \alpha_j \boldsymbol{\beta}^{(j)} \rangle &\geq \alpha_j \left(l(\boldsymbol{\beta}^{(j)}) + \frac{m_1}{2} |\boldsymbol{\beta}_j^{(j)}|^2 \right) \\ \Rightarrow \langle \nabla l(\mathbf{0}), \bar{\boldsymbol{\beta}} \rangle &\geq \sum_{j \in S} \alpha_j l(\boldsymbol{\beta}^{(j)}) + \alpha_j \frac{m_1}{2} |\boldsymbol{\beta}_j^{(j)}|^2. \end{aligned}$$

Combining (9) with (10), and setting

$$\alpha_j = \frac{m_1}{2M_k} + \frac{l(\boldsymbol{\beta}^{(j)})}{M_k |\boldsymbol{\beta}^{(j)}|^2},$$

we have

$$(11) \quad l(\bar{\boldsymbol{\beta}}) \geq \sum_{j \in \mathbf{S}} \frac{m_1}{2M_k} l(\boldsymbol{\beta}^{(j)}) + \frac{m_1^2}{8M_k} |\boldsymbol{\beta}^{(j)}|^2 + \frac{(l(\boldsymbol{\beta}^{(j)}))^2}{2M_k |\boldsymbol{\beta}^{(j)}|^2}.$$

Now applying Definition 3 with $\mathbf{x} = \boldsymbol{\beta}^{(j)}$ and $\mathbf{y} = \mathbf{0}$,

$$(12) \quad \frac{M_1}{2} |\boldsymbol{\beta}^{(j)}|^2 \geq l(\boldsymbol{\beta}^{(j)}) \geq \frac{m_1}{2} |\boldsymbol{\beta}^{(j)}|^2.$$

Combining (11) and (12), we have

$$\begin{aligned} l(\bar{\boldsymbol{\beta}}) &\geq \sum_{j \in \mathbf{S}} \frac{m_1}{2M_k} l(\boldsymbol{\beta}^{(j)}) + \frac{m_1^2}{4M_k M_1} l(\boldsymbol{\beta}^{(j)}) + \frac{m_1}{4M_k} l(\boldsymbol{\beta}^{(j)}) \\ &= \sum_{j \in \mathbf{S}} \left(\frac{3m_1}{4M_k} + \frac{m_1^2}{4M_k M_1} \right) l(\boldsymbol{\beta}^{(j)}). \end{aligned}$$

Since $l(\boldsymbol{\beta}^{(\mathbf{S})})$ optimizes l over all vectors with support in \mathbf{S} ,

$$\begin{aligned} f(\mathbf{S}) = l(\boldsymbol{\beta}^{(\mathbf{S})}) &\geq l(\bar{\boldsymbol{\beta}}) \geq \frac{m_1}{4M_k} \left(3 + \frac{m_1}{M_1} \right) \sum_{j \in \mathbf{S}} l(\boldsymbol{\beta}^{(j)}) \\ &= \frac{m_1}{4M_k} \left(3 + \frac{m_1}{M_1} \right) \sum_{j=1}^k f(j). \end{aligned}$$

□

5.3. Proof of Theorem 2.

PROOF. Let \mathbf{S} be the set of size k selected by the Oblivious algorithm and \mathbf{S}^* be the optimal set of size k corresponding to values f^{OBL} and f^{OPT} . By definition, $\sum_{j \in \mathbf{S}} f(j) \geq \sum_{j \in \mathbf{S}^*} f(j)$. Letting $C = \max\{\frac{1}{k}, \frac{3m}{4M} + \frac{m^2}{4M^2}\}$ and combining Lemma 1 with Theorem 1,

$$\begin{aligned} f^{\text{OBL}} = f(\mathbf{S}) &\geq C \sum_{j \in \mathbf{S}} f(j) \\ &\geq C \sum_{j \in \mathbf{S}^*} f(j) \geq C \gamma_{\emptyset, k} f(\mathbf{S}^*) \geq C \left(\frac{m_k}{M_1} \right) f(\mathbf{S}^*) \\ &= C \left(\frac{m_k}{M_1} \right) f^{\text{OPT}}. \end{aligned}$$

□

5.4. Proof of Theorem 3.

PROOF. Let $l(\cdot)$ be the log-likelihood function and let S_i^G be the set selected by the Forward Stepwise algorithm at iteration i . Define $A(i)$ as the incremental greedy gain $f(S_i^G) - f(S_{i-1}^G)$ with $A(0) = 0$. Denote the remainder set at iteration i as $S_i^R = S^* \setminus S_i^G$, and define $B(i) = f(S^*) - f(S_i^G)$, the incremental gain from adding the optimal set. Lemma 2 relates these two quantities.

LEMMA 2. At iteration i , the incremental gain from selecting the next greedy item is related to the incremental gain from adding the rest of the optimal set S^* by the following:

$$A(i + 1) \geq \frac{\gamma_{S_i^G, k}}{k} B(i).$$

PROOF. Let $S = S_i^G$ be the set selected by the greedy algorithm at iteration i , S^* be the optimal feature set on k variables, and S^R be the remainder set $S^* \setminus S$. S^R is a subset of the candidate variables available to the greedy algorithm at iteration $i + 1$. Using Definition 2 and the fact that $k \geq |S^R|$,

$$\begin{aligned} kA(i + 1) &\geq |S^R|A(i + 1) \geq |S^R| \max_{j \in S^R} f(S \cup j) - f(S) \\ &\geq \sum_{j \in S^R} [f(S \cup j) - f(S)] \\ &\geq \gamma_{S, |S^R|} (f(S \cup S^R) - f(S)) \geq \gamma_{S, k} B(i), \end{aligned}$$

where the last inequality follows from the fact that $S \cup S^R \supseteq S^*$. \square

Given Theorem 1 and Lemma 2, the rest of the proof follows the standard approximation bound for maximizing a normalized, monotone submodular function (refer to [41] or the survey [29]). Next, observe that $A(i + 1) = B(i) - B(i + 1)$. Combining this with Lemma 2 and letting $C = \gamma_{S_k^G, k}/k$, we have the following inequality:

$$B(i + 1) \leq (1 - C)B(i),$$

which implies

$$B(i) \leq (1 - C)^i B(0),$$

for all iterations $1 \leq i \leq k$. Setting $i = k$ and substituting $B(k) = f^{\text{OPT}} - f^{\text{FS}}$ and $B(0) = f^{\text{OPT}}$,

$$\begin{aligned} f^{\text{OPT}} - f^{\text{FS}} &\leq (1 - C)^k f^{\text{OPT}} \\ \Rightarrow f^{\text{FS}} &\geq f^{\text{OPT}} [1 - (1 - C)^k] \geq f^{\text{OPT}} (1 - e^{-\gamma_{S_k^G, k}}). \end{aligned}$$

The claim follows from applying Theorem 1. \square

5.5. Proof of Theorem 4.

PROOF. First, we prove the following lemma which bounds the ratio of the objective between optimal sets S_k and S_{k-1} in terms of their smoothness and convexity parameters.

LEMMA 3. Let S_k be the optimal subset of size k , and let m be the restricted strong concavity parameter on Ω_k . Let k' satisfy $M'/m < k' < k$, where M' is the restricted smoothness parameter of $l(\cdot)$ on $\tilde{\Omega}_k$. Then for large enough k ,

$$l(\beta^{(S_{k'})}) \geq l(\beta^{(S_k)}) \Theta \left(\left(\frac{k'}{k} \right)^{M'/m} \right) \Rightarrow l(\beta^{(S_{k/2})}) \geq l(\beta^{(S_k)}) \Theta(2^{-M'/m}).$$

PROOF. Let j be the index that minimizes $|\beta_j^{(S_k)}|^2$. By M' -smoothness on $\tilde{\Omega}_k$ and the fact that the min is smaller than the average,

$$\begin{aligned} l(\beta^{(S_{k-1})}) &\geq l(\beta_{S_k \setminus \{j\}}^{(S_k)}) \\ &\geq l(\beta^{(S_k)}) + \langle \nabla l(\beta^{(S_k)}), \beta_{S_k \setminus \{j\}}^{(S_k)} - \beta^{(S_k)} \rangle - \frac{M'}{2} \|\beta_{S_k \setminus \{j\}}^{(S_k)} - \beta^{(S_k)}\|_2^2 \\ &= l(\beta^{(S_k)}) - \frac{M'}{2} |\beta_j^{(S_k)}|^2. \end{aligned}$$

This implies

$$\frac{l(\beta^{(S_{k-1})})}{l(\beta^{(S_k)})} \geq 1 - \frac{M' \|\beta^{(S_k)}\|_2^2}{2kl(\beta^{(S_k)})}.$$

Assuming that $l(\beta^{(\emptyset)}) = 0$ and using m -strong concavity on Ω_k ,

$$\begin{aligned} l(\beta^{(\emptyset)}) - l(\beta^{(S_k)}) &\leq -\frac{m}{2} \|\beta^{(S_k)} - \beta^{(\emptyset)}\|_2^2 \\ \Rightarrow -\frac{\|\beta^{(S_k)}\|_2^2}{l(\beta^{(S_k)})} &\geq -\frac{2}{m} \\ \Rightarrow \frac{l(\beta^{(S_{k-1})})}{l(\beta^{(S_k)})} &\geq 1 - \frac{M'}{km}. \end{aligned}$$

Then applying iteratively for M'/m constant, k large, and $M'/m < k' < k$, as in [13] we have

$$l(\beta^{(S_{k'})}) \geq l(\beta^{(S_k)}) \prod_{j=k'+1}^k \left(1 - \frac{M'}{jm} \right) = l(\beta^{(S_k)}) \Theta \left(\left(\frac{k'}{k} \right)^{M'/m} \right). \quad \square$$

Observe that the assumptions of Lemma 3 are satisfied. Combining with Theorem 3,

$$\begin{aligned}
 l(\boldsymbol{\beta}^{(\mathbf{S}_k^G)}) &\geq l(\boldsymbol{\beta}^{(\mathbf{S}_{k/2}^G)}) \geq l(\boldsymbol{\beta}^{(\mathbf{S}_{k/2})})(1 - e^{-\gamma_{\mathbf{S}_{k/2}^G, k/2}^G}) \\
 &\geq l(\boldsymbol{\beta}^{(\mathbf{S}_k)}) \Theta(2^{-M'/m})(1 - e^{-\gamma_{\mathbf{S}_{k/2}^G, k/2}^G}) \\
 \Rightarrow l(\boldsymbol{\beta}^{(\mathbf{S}_k^G)}) &\geq l(\boldsymbol{\beta}^{(\mathbf{S}_k)}) \Theta(2^{-M'/m'})(1 - e^{-m'/M'}). \quad \square
 \end{aligned}$$

5.6. Proof of Theorem 5.

PROOF. The key idea at each step i is to lower bound the incremental gain from the index chosen by OMP. This is similar to the proof of Theorem 3, as well as [26] in which a matrix completion objective is considered. Let $\mathbf{S} = \mathbf{S}_i^P$ be the set chosen by OMP up to iteration i . Given \mathbf{S} , let v be the index that would be selected by running one additional step of OMP. Define $D(i + 1) = f(\mathbf{S}_{i+1}^P) - f(\mathbf{S}) = l(\boldsymbol{\beta}^{(\mathbf{S} \cup \{v\})}) - l(\boldsymbol{\beta}^{(\mathbf{S})})$, and define $\tilde{B}(i) = f(\mathbf{S}^*) - f(\mathbf{S})$.

LEMMA 4. At iteration i , the incremental gain from selecting the next item via OMP is related to the incremental gain from adding the rest of the optimal set \mathbf{S}^* by the following:

$$D(i + 1) \geq \frac{m_{i+k}}{k\tilde{M}_{i+1}} \tilde{B}(i).$$

PROOF. We begin similar to the proof of Theorem 1. Let $M = \tilde{M}_{i+1}$, $m = m_{i+k}$, and \mathbf{e}_v be the unit vector with one at coordinate v . By Definition 3 with $\mathbf{x} = \boldsymbol{\beta}^{(\mathbf{S})}$ and $\mathbf{y} = \boldsymbol{\beta}^{(\mathbf{S})} + \alpha \mathbf{e}_v$ for any scalar α ,

$$\begin{aligned}
 D(i + 1) &\geq l(\mathbf{y}) - l(\mathbf{x}) \geq \langle \nabla l(\boldsymbol{\beta}^{(\mathbf{S})}), \alpha \mathbf{e}_v \rangle - \frac{M}{2} \alpha^2 \\
 &= \alpha \|\nabla l(\boldsymbol{\beta}^{(\mathbf{S})})\|_\infty - \frac{M}{2} \alpha^2,
 \end{aligned}$$

since OMP chooses the coordinate which maximizes the gradient. Substituting

$$\alpha = \frac{\|\nabla l(\boldsymbol{\beta}^{(\mathbf{S})})\|_\infty}{M},$$

we have

$$D(i + 1) \geq \frac{1}{2M} \|\nabla l(\boldsymbol{\beta}^{(\mathbf{S})})\|_\infty^2.$$

Let $\mathbf{S}^R = \mathbf{S}^* \setminus \mathbf{S}$. Since $|\mathbf{S}^R| \leq k$,

$$D(i + 1) \geq \frac{1}{2kM} \sum_{j \in \mathbf{S}^R} \langle \nabla l(\boldsymbol{\beta}^{(\mathbf{S})}), \mathbf{e}_j \rangle^2 = \frac{1}{2kM} \|\nabla l(\boldsymbol{\beta}^{(\mathbf{S})})_{\mathbf{S}^R}\|_2^2.$$

Substituting (8) into the above and noting that $\mathbf{S} \cup \mathbf{S}^R \supseteq \mathbf{S}^*$, we have

$$D(i + 1) \geq \frac{m}{kM} (l(\boldsymbol{\beta}^{(\mathbf{S} \cup \mathbf{S}^R)}) - l(\boldsymbol{\beta}^{(\mathbf{S})})) \geq \frac{m}{kM} \tilde{B}(i). \quad \square$$

Given Lemma 4, the rest of the proof follows that of Theorem 3. \square

5.7. Proof of Theorem 6.

PROOF. Let $C = C_{s,r}$ and $\Delta = \widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^s$, which is at most an $(s + r)$ -sparse vector. Recall that by the definition of restricted strong concavity on Ω_{s+r} we have

$$(13) \quad l(\widehat{\boldsymbol{\beta}}^r) - l(\boldsymbol{\beta}^s) - \langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle \leq \frac{-m_{s+r}}{2} \|\Delta\|_2^2.$$

Furthermore, simple calculations show that

$$(14) \quad l(\widehat{\boldsymbol{\beta}}^r) - l(\boldsymbol{\beta}^s) \geq (1 - C)[l(\mathbf{0}) - l(\boldsymbol{\beta}^s)].$$

Subtracting $\langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle$ from both sides of (14), we have

$$l(\widehat{\boldsymbol{\beta}}^r) - l(\boldsymbol{\beta}^s) - \langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle \geq -\langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle + (1 - C)[l(\mathbf{0}) - l(\boldsymbol{\beta}^s)].$$

Applying (13) yields

$$\frac{-m_{s+r}}{2} \|\Delta\|_2^2 \geq -\langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle + (1 - C)[l(\mathbf{0}) - l(\boldsymbol{\beta}^s)].$$

Next, note that

$$-\langle \nabla l(\boldsymbol{\beta}^s), \Delta \rangle \geq -\|\nabla l(\boldsymbol{\beta}^s)\|_{2,s+r} \|\Delta\|_2.$$

Thus,

$$\frac{-m_{s+r}}{2} \|\Delta\|_2^2 \geq -\|\nabla l(\boldsymbol{\beta}^s)\|_{2,k} \|\Delta\|_2 + (1 - C)[l(\mathbf{0}) - l(\boldsymbol{\beta}^s)].$$

Recalling that for any positive numbers $2ab \leq ca^2 + b^2/c$ and flipping the above inequality,

$$\frac{m_{s+r}}{2} \|\Delta\|_2^2 \leq \frac{\|\nabla l(\boldsymbol{\beta}^s)\|_{2,s+r}^2}{m_{s+r}} + \frac{m_{s+r} \|\Delta\|_2^2}{4} + (1 - C)[l(\boldsymbol{\beta}^s) - l(\mathbf{0})].$$

Rearranging terms we have the final result. \square

5.8. Proof of Corollary 3.

PROOF. Using a result of [40], we have that

$$\|\nabla l(\boldsymbol{\beta}^s)\|_{2,s+r}^2 \leq (s + r) \|\nabla l(\boldsymbol{\beta}^s)\|_\infty^2 \leq \frac{(s + r)\sigma^2 \log p}{n},$$

with probability at least $1 - 1/p$.

The minimum eigenvalue of the matrix Σ is 1, while the maximum s -sparse eigenvalue behaves like $1 + s$. Hence, an RIP type condition will not hold. However, in our setting, we simply require a bound on M_1 . It can be shown using tail bounds for χ^2 -random variables that with high probability $M_1 \leq 4$. Letting $\rho(\Sigma)^2 = \max_i \Sigma_{ii}$, and using a result by Raskutti et al. [44], we have that for all $\mathbf{v} \in \mathbb{R}^p$,

$$\begin{aligned} \frac{\|\mathbf{X}\mathbf{v}\|_2}{\sqrt{n}} &\geq \frac{1}{4} \|\Sigma^{1/2}\mathbf{v}\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log p}{n}}\|\mathbf{v}\|_1 \\ &\geq \left(\left(1 - \frac{1}{c}\right)\frac{\lambda_{\min}(\Sigma)}{16} + (1 - c)81\rho(\Sigma)^2\frac{\log p}{n}(s + r) \right) \|\mathbf{v}\|_2^2 \\ \Rightarrow m_{s+r} &\geq \min_{\substack{\mathbf{v}: \|\mathbf{v}\|_2=1, \\ \|\mathbf{v}\|_0 \leq s+r}} \frac{\|\mathbf{X}\mathbf{v}\|_2^2}{n} \geq \frac{1}{32} - \frac{162(s + r)\log p}{n}, \end{aligned}$$

with high probability. Therefore, $\gamma \geq \frac{1}{128} - \frac{81(s+r)\log p}{2n}$ and with probability at least $1 - p^{-\Omega(1)} - e^{-\Omega(n)}$,

$$\|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2^2 \leq \frac{4}{m_{s+r}} \frac{(s + r)\sigma^2 \log p}{n} + \frac{8(s + 1)}{m_{s+r}}(1 - C_{s,r}),$$

where we have used the fact that $l(\boldsymbol{\beta}^*) - l(\mathbf{0}) \leq \lambda_{\max}(\widehat{\Sigma}_s) \leq 2(s + 1)$ with high probability. Note that using arguments from [32, 39] we can apply the above results to the setting of generalized linear models.

Now let $(s + r)\sigma^2 \log p = o(n)$ and $r = \Omega(s \log n)$. Combined with Corollary 1, this implies that $\|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2^2 = n^{-\Omega(1)}$ with probability $1 - p^{-\Omega(1)} - e^{-\Omega(n)}$. \square

6. Experiments. Next, we evaluate the performance of our greedy algorithms with feature selection experiments on simulated and real-world datasets. A bias term β_0 is added to the regression by augmenting the design matrix with a column of ones.

The data. A synthetic experiment was conducted as follows: first each row of a 600×200 design matrix \mathbf{X} is generated independently according to a first order AR process ($\alpha = 0.3$ and noise variance $\sigma^2 = 5$). This ensures that the features are heavily correlated with each other. Bernoulli ± 1 (i.e., Rademacher) random variables are placed on 50 random indices to form the true support $\bar{\boldsymbol{\beta}}^k$, and scaled such that $\|\boldsymbol{\beta}\|_2^2 = 5$. Then responses \mathbf{y} are computed via a logistic model. We also conduct an experiment on a subset of the RCV1 Binary text classification dataset [30]. 10,000 training and test samples are used in 47,236 dimensions. Since there is no ground truth, a logistic regression is fit using a subset of at most 700 features.

Algorithms and metrics. The Oblivious, Forward Stepwise (FS), and OMP algorithms were implemented using a logistic log-likelihood function given \mathbf{X} and \mathbf{y} (see Appendix B). We implemented 3 additional algorithms. *Lasso* fits a logistic regression model with ℓ_1 regularization. *Lasso-Pipeline* recovers the sparse support using Lasso and then fits regression coefficients on this support with a separate, unregularized model. The regularization parameter was swept to achieve outputs with varying sparsity levels. *Forward Backward* (FoBa) [52] first runs FS at each step and then drops any features if doing so would decrease the objective by less than half the latest marginal gain.

Our main metric for each algorithm is the normalized objective function $l(\widehat{\boldsymbol{\beta}}^s) - l(\mathbf{0})$ for the output sparsity $s \in \{1, \dots, 70\}$. We also compare the sets $\text{supp}(\widehat{\boldsymbol{\beta}}^s)$ and $\text{supp}(\bar{\boldsymbol{\beta}}^k)$ using area under ROC and percent of true support recovered. Finally, we measure generalization accuracy by drawing additional observations (\mathbf{x}_i, y_i) from the same distribution as the training data.

Results. Figure 1 shows the results of our synthetic experiment averaged over 20 runs. For all metrics, Oblivious performs worse than OMP which is slightly worse than FS and FoBa. This matches intuition and the series of bounds in Section 3. We also see that the Lasso-Pipeline performs noticeably worse than all algorithms except Oblivious and Lasso. This suggests that greedy feature selection degrades more gracefully than Lasso in the case of correlated features.

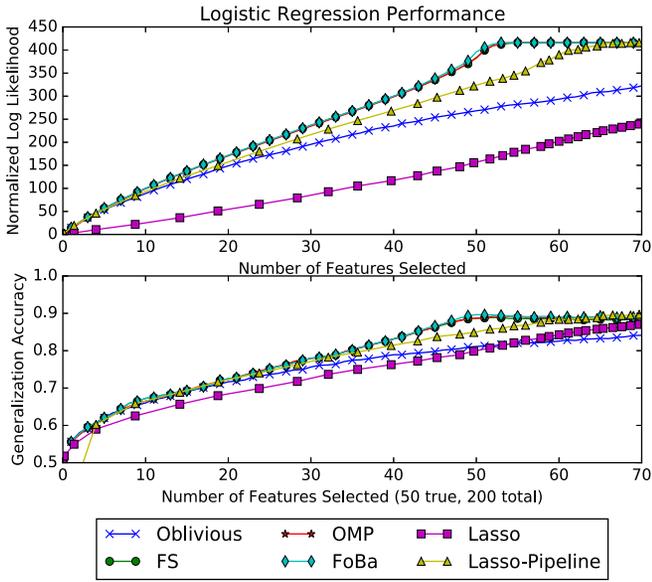
Figure 2 shows similar results for the high-dimensional RCV1 Binary dataset. Due to their large running time complexity, FS and FoBa were omitted. While all algorithms have roughly the same generalization accuracy using 300 features, OMP has the largest log-likelihood.

7. Conclusions. We have extended the results of [13] and shown that functions satisfying RSC also satisfy a relaxed form of submodularity that can be used to analyze the performance of greedy algorithms compared to the best sparse solution. Experimental results confirm that greedy feature selection outperforms regularized approaches in a nonlinear regression model. Directions for future work include similar analysis for other greedy algorithms that incorporate group sparsity [40] or thresholding, and applications beyond sparse regression. Bounds on dictionary selection (analogous to those in [13]) also apply to general likelihood functions satisfying RSC and RSM.

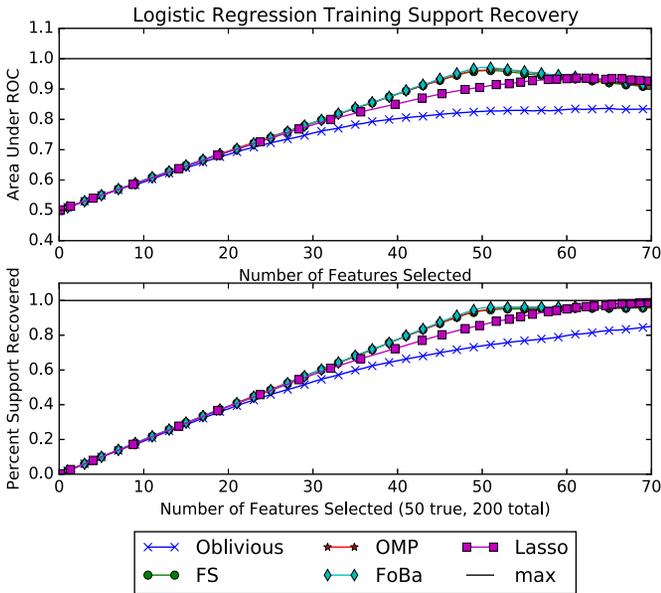
APPENDIX A: MOTIVATING EXAMPLE (LINEAR REGRESSION)

To show the impact of submodularity, we construct a linear regression example. Even in $p = 3$ dimensions, the greedy forward selection algorithm's output can be arbitrarily off from the optimal R^2 . Consider the following variables:

$$\mathbf{y} = [1 \quad 0 \quad 0]^T,$$



(a)



(b)

FIG. 1. Synthetic dataset— $\alpha = 0.3$, $n = 600$ training and test samples, $p = 200$ dimensions with true support on 50 features, averaged over 20 runs. (a) The greedy algorithms perform better than Lasso and Oblivious algorithms, but beyond 50 steps they overfit to noise in the training data. While Lasso outperforms Oblivious in support recovery (b), its regression suffers from regularization bias.

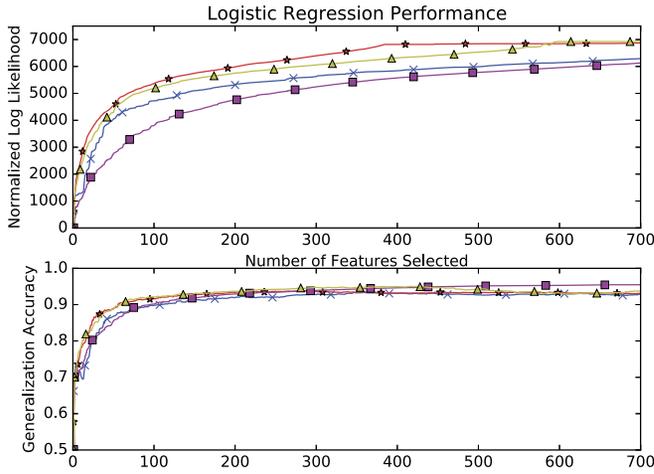


FIG. 2. *RCV1 binary dataset*— $n = 10,000$, $p = 47,236$. *OMP outperforms Lasso-Pipeline.*

$$\begin{aligned} \mathbf{x}_1 &= [0 \quad 1 \quad 0]^T, \\ \mathbf{x}_2 &= [z \quad \sqrt{1-z^2} \quad 0]^T, \\ \mathbf{x}_3 &= [2z \quad 0 \quad \sqrt{1-4z^2}]^T. \end{aligned}$$

All variables have unit norm and we wish to choose the 2-subset of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ that best estimates \mathbf{y} . Since $R_1^2 = 0$, $R_2^2 = z^2$, and $R_3^2 = 4z^2$, \mathbf{x}_3 will be selected first ($S_1^G = \{3\}$) if $z > 0$. \mathbf{x}_2 will be chosen next ($S_2^G = \{3, 2\}$), and solving for R^2 for this pair,

$$\begin{aligned} R_{3,2}^2 &= (\mathbf{y}^T \mathbf{X}_{3,2})(\mathbf{X}_{3,2}^T \mathbf{X}_{3,2})^{-1}(\mathbf{X}_{3,2}^T \mathbf{y}) \\ &= \frac{1}{1-4z^4} [2z \quad z] \begin{bmatrix} 1 & -2z^2 \\ -2z^2 & 1 \end{bmatrix} \begin{bmatrix} 2z \\ z \end{bmatrix} = \frac{5z^2 - 8z^4}{1-4z^4}, \end{aligned}$$

which goes to zero as $z \rightarrow 0^+$. However, $\mathbf{y} = -\frac{\sqrt{1-z^2}}{z}\mathbf{x}_1 + \frac{1}{z}\mathbf{x}_2$ which makes $R_{1,2}^2 = 1$ for the optimal set $\{\mathbf{x}_1, \mathbf{x}_2\}$ ($S_2 = \{1, 2\}$).

APPENDIX B: GREEDY SELECTION FOR GLMS

In this section, we state guarantees for feature selection in sparse generalized linear regression using the framework developed in Section 3. For introducing sparsity, a relevant regularizer (such as ℓ_1) is often used. An alternative to regularization is to apply Algorithms 1–3 to the log-likelihood function. To use guarantees presented in Section 3, we derive sample complexity conditions on the design matrix \mathbf{X} that are sufficient to bound the submodularity ratio $\gamma_{U,k}$ with high probability.

Recall that Theorems 2–4 require strong convexity and bounded smoothness on a sparse support. While in general GLMs are not strongly convex, nor do they have bounded smoothness, it can be shown that on the *restricted* set of sparse supports under some mild restrictions on the design matrix, they possess both these traits. Moreover, note that some regularizers such as ℓ_2 that are widely applied for GLMs automatically imply strong convexity. We present the analysis for both regularized and unregularized regression.

For continuity, we reintroduce some notation here. We represent the data as (\mathbf{x}_i, y_i) , where $\forall i, \mathbf{x}_i \in \mathbb{R}^p$ are features, and $y_i \in \mathbb{R}$ represents the response. The log conditional can be written in its *canonical* form as [14]:

$$(15) \quad \log p(y|\mathbf{x}; \boldsymbol{\beta}) = h^{-1}(\tau)y\boldsymbol{\beta}^\top \mathbf{x} - Z(\boldsymbol{\beta}, \mathbf{x}) + g(y, \tau),$$

where $Z(\cdot)$ is the log partition function, and $\boldsymbol{\beta}, \tau$ are the parameters (τ is also called the dispersion parameter). For n observations, we can write the log-likelihood as

$$(16) \quad l_{\text{GLM}}(\boldsymbol{\beta}) := \sum_{i=1}^n \log p(y_i|\mathbf{x}_i; \boldsymbol{\beta}).$$

The parameters of this distribution can then be learned by maximizing the log-likelihood. Recall that equivalently we can minimize its negative. Also, typically a regularization term is added for stability and identifiability. The *loss* function g to minimize for learning can be written as

$$(17) \quad g(\boldsymbol{\beta}) := -l_{\text{GLM}}(\boldsymbol{\beta}) + \eta \|\boldsymbol{\beta}\|_2^2.$$

Similar to the (4), we can define a normalized set function $f_{\text{GLM1}}(\cdot)$ associated with $g(\cdot)$ as

$$\max_{|\mathbf{S}| \leq k} f_{\text{GLM1}}(\mathbf{S}) \Leftrightarrow \min_{\substack{\boldsymbol{\beta}: \boldsymbol{\beta}_{\mathbf{S}^c} = 0 \\ |\mathbf{S}| \leq k}} g(\boldsymbol{\beta}).$$

Note that increasing the support set of $\boldsymbol{\beta}$ does not decrease the log-likelihood, so our set function $f_{\text{GLM1}}(\cdot)$ is indeed monotone. Further, note that for normalizing the set function, we can use $f_{\text{GLM1}}(\emptyset) = g(\mathbf{0})$.

The Hessian of g at any point $\boldsymbol{\beta}$ can be written as

$$(18) \quad \mathbf{H}(\boldsymbol{\beta}) := \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \mathbf{X}^\top \mathbf{D} \mathbf{X} + \eta \mathbf{I},$$

where \mathbf{D} is a diagonal matrix with $\mathbf{D}_{ii} = h^{-1}(\tau)Z''(\boldsymbol{\beta}, \mathbf{x}_i)$. Next, we state assumptions required for the sample complexity bounds. We assume that \mathbf{D}_{ii} is upper bounded by s at any value of the domain of $\boldsymbol{\beta}$. This implies that $\eta \mathbf{I} \preceq \mathbf{H}(\boldsymbol{\beta}) \preceq s \mathbf{X}^\top \mathbf{X} + \eta \mathbf{I}$ for all $\boldsymbol{\beta}$.

Let $\mathbf{T} \subset [p]$ so that $|\mathbf{T}| \leq r$. For any vector \mathbf{v} , define $\mathbf{v}_\mathbf{T}$ be the vector formed by replacing all indices in $[p] \setminus \mathbf{T}$ of \mathbf{v} by 0. Let $\mathcal{P}_\mathbf{T}$ be the operator that achieves this, that is, $\mathcal{P}_\mathbf{T} \mathbf{v} = \mathbf{v}_\mathbf{T}$.

ASSUMPTION 1. We make the following assumptions for Proposition 2. Let the rows of \mathbf{X} be generated i.i.d. from some underlying distribution, so that $\mathbb{E}[\mathbf{xx}^\top] = \mathbf{C}$. For all $\mathsf{T} \subset [p]$, $|\mathsf{T}| \leq r$:

1. $\|\mathbf{x}_{\mathsf{T}}\|_2 \leq R$, and
2. None of the matrices $\mathcal{P}_{\mathsf{T}}^\top \mathbf{C} \mathcal{P}_{\mathsf{T}}$ are the zero matrix.

Further, define $\theta_{\mathsf{T}} := \lambda_{\max}(\mathcal{P}_{\mathsf{T}}^\top \mathbf{C} \mathcal{P}_{\mathsf{T}})$, let $\bar{\theta} = \max_{\mathsf{T} \subset [p], |\mathsf{T}| \leq r} \theta_{\mathsf{T}}$, $\tilde{\theta} = \min_{\mathsf{T} \subset [p], |\mathsf{T}| \leq r} \theta_{\mathsf{T}}$.

PROPOSITION 2 (From [3]). *With Assumption 1, for $\delta \in (0, 1)$ and $n > \frac{R(\log r + r(1 + \log \frac{p}{k} - \log \delta))}{\tilde{\theta}(1 + \varepsilon) \log(1 + \varepsilon) - \varepsilon}$, $\lambda_{\max}(\mathcal{P}_{\mathsf{T}}^\top \mathbf{X}^\top \mathbf{X} \mathcal{P}_{\mathsf{T}}) \leq (1 + \varepsilon)\bar{\theta}$ with probability $(1 - \delta)$.*

COROLLARY 4 (Regularized GLM sample complexity). *Under Assumption 1, for $\delta \in (0, 1)$ and $n > \frac{R(\log r + r(1 + \log \frac{p}{k} - \log \delta))}{\tilde{\theta}(1 + \varepsilon) \log(1 + \varepsilon) - \varepsilon}$, with probability $(1 - \delta)$, the submodularity ratio of f_{GLM1} , $\gamma_{\mathbf{U}, r} \geq \eta / (\eta + s(1 + \varepsilon)\bar{\theta})$.*

PROOF. Note that f_{GLM1} is η -strongly convex. Further, from (18), and Proposition 2, it is $(\eta + s(1 + \varepsilon)\bar{\theta})$ -smooth with probability $(1 - \delta)$. The result now follows from Theorem 1. \square

REMARK 9. The above discussion can also be motivated by the restricted stability property of the Hessian of the loss function. Define:

$$\mathbf{A}_r(\boldsymbol{\beta}) := \max\{\mathbf{v}^\top \mathbf{H}(\boldsymbol{\beta}) \mathbf{v} \mid |\text{supp}(\boldsymbol{\beta}) \cup \text{supp}(\mathbf{v})| \leq r, \|\mathbf{v}\|_2 = 1\},$$

$$\mathbf{B}_r(\boldsymbol{\beta}) := \min\{\mathbf{v}^\top \mathbf{H}(\boldsymbol{\beta}) \mathbf{v} \mid |\text{supp}(\boldsymbol{\beta}) \cup \text{supp}(\mathbf{v})| \leq r, \|\mathbf{v}\|_2 = 1\}.$$

The Hessian is said to be μ_r -SRH (Stable Restricted Hessian) [3] if, $\mu_r \geq \frac{\mathbf{A}_r}{\mathbf{B}_r}$. It is straightforward to see that $\gamma_{\mathbf{U}, r} \geq 1/\mu_r$.

B.1. Restricted strong convexity of GLMs. Under stronger assumptions on the design matrix, \mathbf{X} , it is possible to have $\gamma > 0$ even when $\eta = 0$ (i.e., the loss is unregularized) in Corollary 4. In the following discussion, we assume $\eta = 0$ to avoid clutter, but the discussion can be readily extended to the case when $\eta > 0$. Similar to the case study of regular GLMs, we consider the GLM loss as the negative log-likelihood, defined as

$$(19) \quad h(\boldsymbol{\beta}) = -l_{\text{GLM}}(\boldsymbol{\beta}),$$

and provide sample complexity bounds for weak submodularity to hold for the associated set function:

$$\max_{|\mathsf{S}| \leq k} f_{\text{GLM2}}(\mathsf{S}) \Leftrightarrow \min_{\substack{\boldsymbol{\beta}: \boldsymbol{\beta}_{\mathsf{S}^c} = 0 \\ |\mathsf{S}| \leq k}} h(\boldsymbol{\beta}).$$

Restricted strong convexity for GLMs was studied by Negahban et al. [40] with the restricted sets being the neighborhood of the true optimum of a convex loss function. This allows GLMs to satisfy RSC on sparse models with support on the true optimum. Loh and Wainwright [32] extended the RSC conditions to hold uniformly for all k -sparse models for GLMs. We present the requisite results here.

Recall that $D_h(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$. For brevity, we drop the subscript, and use $D(\mathbf{x}, \mathbf{y})$. Also recall that $h(\cdot)$ is m -strongly convex on a set \mathcal{S} if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}, D(\mathbf{x}, \mathbf{y}) \geq \frac{m}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$, and that $h(\cdot)$ is M -smooth on the same domain if $D(\mathbf{x}, \mathbf{y}) \leq \frac{M}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$. Let $\mathbb{B}_m(r)$ represent an m -norm ball of radius r .

ASSUMPTION 2. The design matrix \mathbf{X} consists of samples drawn i.i.d. from a sub-Gaussian distribution with parameter $\sigma_{\mathbf{x}}^2$, and covariance matrix Σ .

THEOREM 7 (From [32]). *If Assumption 2 is true, there exists a constant α_q depending on the GLM family, and on $\sigma_{\mathbf{x}}^2, \Sigma, q$ such that for all vectors $\mathbf{y} \in \mathbb{B}_2(3) \cap \mathbb{B}_1(q)$ for a constant q s.t. $q\sqrt{\frac{\log p}{n}} \lesssim 1$, so that with probability $1 - c_1 \exp(-c_2 n)$,*

$$D(\mathbf{x}, \mathbf{y}) \geq \begin{cases} \frac{\alpha_q}{2} \|\Delta\|_2^2 - \frac{c^2 \sigma_{\mathbf{x}} \log p}{2\alpha_q n} \|\Delta\|_1^2 & \text{if } \|\Delta\|_2 \leq 3, \\ \frac{3\alpha_q}{2} \|\Delta\|_2 - 3c\sigma_{\mathbf{x}} \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{otherwise,} \end{cases}$$

where $\Delta = \mathbf{x} - \mathbf{y}$. Similarly, $D(\mathbf{x}, \mathbf{y})$ can be upper bounded. Recall that s is such that $\max_i \mathbf{D}_{ii} \leq s$ for \mathbf{D} as used in (18). Then, with probability $1 - c_1 \exp(-c_2 n)$,

$$D(\mathbf{x}, \mathbf{y}) \leq s \lambda_{\max}(\Sigma) \left(\frac{3}{2} \|\Delta\|_2^2 + \frac{\log p}{n} \|\Delta\|_1^2 \right).$$

Theorem 7 can be applied to r -sparse sets to get the sample complexity for strong convexity. We further assume that $\text{supp}(\mathbf{x}) \subset \text{supp}(\mathbf{y})$ or $\text{supp}(\mathbf{y}) \subset \text{supp}(\mathbf{x})$ which is not restrictive for our analysis in Section 3. This implies if \mathbf{x}, \mathbf{y} are r -sparse, $|\text{supp}(\Delta)| \leq r$.

COROLLARY 5 (Sample complexity sub-Gaussian design). *Under Assumption 2, for $n > \frac{c^2 \sigma_{\mathbf{x}}}{\alpha_q^2} (r + |\mathcal{U}|) \log p$, the submodularity ratio for $f_{\text{GLM2}}(\cdot), \gamma_{\mathcal{U}, r} \geq m'/M'$, where $m' = (\alpha_q - \frac{c^2 \sigma_{\mathbf{x}} (r + |\mathcal{U}|) \log p}{n}) > 0$, and $M' = \lambda_{\max}(\Sigma) (\frac{3}{2} + \frac{(r + |\mathcal{U}|) \log p}{n})$.*

PROOF. Since $\|\Delta\|_1 \leq \sqrt{(r + |\mathcal{U}|)} \|\Delta\|_2$, from Theorem 7, $D(\mathbf{x}, \mathbf{y}) \geq \frac{1}{2} (\alpha_q - \frac{c^2 \sigma_{\mathbf{x}} (r + |\mathcal{U}|) \log p}{n}) \|\Delta\|_2^2$. The sample complexity bound follows by ensuring the RHS is > 0 . This gives $h(\cdot)$ to be m -strongly convex with $m \geq (\alpha_q - \frac{c^2 \sigma_{\mathbf{x}} (r + |\mathcal{U}|) \log p}{n})$.

Similarly, a corresponding version of restricted smoothness by using the upper bound of $D(\mathbf{x}, \mathbf{y})$ in Theorem 7 and using $\|\Delta\|_1 \leq \sqrt{r + |\mathcal{U}|}\|\Delta\|_2$. The expression for the submodularity ratio then follows from Theorem 1. \square

Acknowledgments. We would like to thank the Associate Editor and two anonymous reviewers for their helpful comments.

REFERENCES

- [1] ALTSCHULER, J., BHASKARA, A., FU, G. T., MIRROKNI, V., ROSTAMIZADEH, A. and ZADIMOGHADDAM, M. (2016). Greedy column subset selection: New bounds and distributed algorithms. In *ICML* 2539–2548.
- [2] BACH, F. R. (2013). Learning with submodular functions: A convex optimization perspective. *Found. Trends Mach. Learn.* **6**.
- [3] BAHMANI, S., RAJ, B. and BOUFONOS, P. T. (2013). Greedy sparsity-constrained optimization. *J. Mach. Learn. Res.* **14** 807–841.
- [4] BARBOSA, R. D. P., ENE, A., NGUYEN, H. L. and WARD, J. (2015). The power of randomization: Distributed submodular maximization on massive datasets. In *ICML* 1236–1244.
- [5] BARRON, A. R., COHEN, A., DAHMEN, W. and DEVORE, R. A. (2008). Approximation and learning by greedy algorithms. *Ann. Statist.* **36** 64–94. [MR2387964](#)
- [6] BIAN, A. A., BUHMANN, J. M., KRAUSE, A. and TSCHIATSCHKEK, S. (2017). Guarantees for greedy maximization of non-submodular functions with applications. In *ICML* 498–507.
- [7] BIAN, A. A., MIRZASOLEIMAN, B., BUHMANN, J. M. and KRAUSE, A. (2017). Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *AISTATS* 111–120.
- [8] BLUMENSATH, T. and DAVIES, M. E. (2007). On the difference between orthogonal matching pursuit and orthogonal least squares. Available at <https://eprints.soton.ac.uk/142469/>.
- [9] BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. [MR882001](#)
- [10] BUCHBINDER, N. and FELDMAN, M. (2016). Constrained submodular maximization via a non-symmetric technique. Available at [arXiv:1611.03253](https://arxiv.org/abs/1611.03253).
- [11] CHEN, L., FELDMAN, M. and KARBASI, A. (2017). Weakly submodular maximization beyond cardinality constraints: does randomization help greedy? Available at [arXiv:1707.04347](https://arxiv.org/abs/1707.04347).
- [12] CONFORTI, M. and CORNUÉJOLS, G. (1984). Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the Rado–Edmonds theorem. *Discrete Appl. Math.* **7** 251–274.
- [13] DAS, A. and KEMPE, D. (2011). Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *ICML* 1057–1064.
- [14] DEY, D. K., GHOSH, S. K. and MALLICK, B. K., eds. (2000). *Generalized Linear Models: A Bayesian Perspective*, 1st ed. CRC Press, Boca Raton, FL.
- [15] EGHBALI, R. and FAZEL, M. (2016). Designing smoothing functions for improved worst-case competitive ratio in online optimization. In *NIPS* 3279–3287.
- [16] ELENBERG, E. R., DIMAKIS, A. G., FELDMAN, M. and KARBASI, A. (2017). Streaming weak submodularity: Interpreting neural networks on the fly. In *NIPS* 4047–4057.
- [17] FUJISHIGE, S. (2005). *Submodular Functions and Optimization*. Elsevier Science, Amsterdam.
- [18] HASSANI, H., SOLTANOLKOTABI, M. and KARBASI, A. (2017). Gradient methods for submodular maximization. In *NIPS* 5843–5853.

- [19] HAZAN, E. and KALE, S. (2012). Online submodular minimization. *J. Mach. Learn. Res.* **13** 2903–2922.
- [20] HOI, S. C. H., JIN, R., ZHU, J. and LYU, M. R. (2006). Batch mode active learning and its application to medical image classification. In *ICML* 417–424.
- [21] HOREL, T. and SINGER, Y. (2016). Maximization of approximately submodular functions. In *NIPS* 3045–3053.
- [22] IYER, R. K. and BILMES, J. (2015). Polyhedral aspects of submodularity, convexity and concavity. Available at [arXiv:1506.07329](https://arxiv.org/abs/1506.07329).
- [23] JAIN, P., TEWARI, A. and KAR, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. In *NIPS* 685–693.
- [24] JALALI, A., JOHNSON, C. and RAVIKUMAR, P. (2011). On learning discrete graphical models using greedy methods. In *NIPS* 1935–1943.
- [25] KAKADE, S. M., SHAMIR, O., SRIDHARAN, K. and TEWARI, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS* 381–388.
- [26] KHANNA, R., ELENBERG, E. R., DIMAKIS, A. G., GHOSH, J. and NEGAHBAN, S. (2017). On approximation guarantees for greedy low rank optimization. In *ICML* 1837–1846.
- [27] KHANNA, R., ELENBERG, E. R., DIMAKIS, A. G., NEGAHBAN, S. and GHOSH, J. (2017). Scalable greedy feature selection via weak submodularity. In *AISTATS* 1560–1568.
- [28] KRAUSE, A. and CEVHER, V. (2010). Submodular dictionary selection for sparse representation. In *ICML* 567–574.
- [29] KRAUSE, A. and GOLOVIN, D. (2014). Submodular function maximization. In *Tractability* 71–104. Cambridge Univ. Press, Cambridge. [MR3220900](https://arxiv.org/abs/1302.0090)
- [30] LEWIS, D. D., YANG, Y., ROSE, T. G. and LI, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* **5** 361–397.
- [31] LIU, J., YE, J. and FUJIMAKI, R. (2014). Forward–backward greedy algorithms for general convex smooth functions over a cardinality constraint. In *ICML* 503–511.
- [32] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.
- [33] LOVÁSZ, L. (1983). Submodular functions and convexity. In *Mathematical Programming the State of the Art: Bonn 1982* 235–257. Springer, Berlin.
- [34] LOZANO, A. C., ŚWIRSZCZ, G. and ABE, N. (2011). Group orthogonal matching pursuit for logistic regression. In *AISTATS* 452–460.
- [35] MAZUMDER, R., FREUND, R. and GRIGAS, P. (2017). A new perspective on boosting in linear regression via subgradient optimization and relatives. *Ann. Statist.* **45** 2328–2364. [MR3737894](https://arxiv.org/abs/1703.07442)
- [36] MIRZASOLEIMAN, B., BADANIDIYURU, A., KARBASI, A., VONDRÁK, J. and KRAUSE, A. (2015). Lazier than lazy greedy. In *AAAI* 1812–1818.
- [37] NATARAJAN, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24** 227–234.
- [38] NEEDELL, D. and TROPP, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **3** 301–321.
- [39] NEGAHBAN, S., RAVIKUMAR, P., YU, B. and WAINWRIGHT, M. J. (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Available at [arXiv:1010.2731v1](https://arxiv.org/abs/1010.2731).
- [40] NEGAHBAN, S., RAVIKUMAR, P., YU, B. and WAINWRIGHT, M. J. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. [MR3025133](https://arxiv.org/abs/1205.1333)
- [41] NEMHAUSER, G. L., WOLSEY, L. A. and FISHER, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. *Math. Program.* **14** 265–294.
- [42] PAN, X., JEGELKA, S., GONZALEZ, J. E., BRADLEY, J. K. and JORDAN, M. I. (2014). Parallel double greedy submodular maximization. In *NIPS* 118–126.

- [43] QIAN, C., YU, Y. and ZHOU, Z.-H. (2015). Subset selection by Pareto optimization. In *NIPS* 1774–1782.
- [44] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. [MR2719855](#)
- [45] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771.
- [46] TEWARI, A., RAVIKUMAR, P. and DHILLON, I. S. (2011). Greedy algorithms for structurally constrained high dimensional problems. In *NIPS* 882–890.
- [47] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645.
- [48] WEI, K., LIU, Y., KIRCHHOFF, K. and BILMES, J. (2013). Using document summarization techniques for speech data subset selection. In *NAACL-HLT* 721–726.
- [49] WEI, K., RISHABH, I. and BILMES, J. (2015). Submodularity in data subset selection and active learning. In *ICML* 1954–1963.
- [50] YANG, Z., WANG, Z., LIU, H., ELДАР, Y. C. and ZHANG, T. (2016). Sparse nonlinear regression: Parameter estimation and asymptotic inference. In *ICML* 2472–2481.
- [51] YUAN, X.-T., LI, P. and ZHANG, T. (2014). Gradient hard thresholding pursuit for sparsity-constrained optimization. In *ICML* 127–135.
- [52] ZHANG, T. (2008). Adaptive forward–backward greedy algorithm for sparse learning with linear models. In *NIPS* 1921–1928.
- [53] ZHANG, T. (2011). Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inform. Theory* **57** 6215–6221. [MR2857968](#)

E. R. ELENBERG
R. KHANNA
A. G. DIMAKIS
DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
UNIVERSITY OF TEXAS AT AUSTIN
1 UNIVERSITY STATION
AUSTIN, TEXAS 78712
USA
E-MAIL: elenberg@utexas.edu
rajivak@utexas.edu
dimakis@austin.utexas.edu

S. NEGAHBAN
DEPARTMENT OF STATISTICS AND
DATA SCIENCE
YALE UNIVERSITY
24 HILLHOUSE
NEW HAVEN, CONNECTICUT 06510
USA
E-MAIL: sahand.negahban@yale.edu