Restricted Strong Convexity Implies Weak Submodularity

Ethan R. Elenberg[1], Rajiv Khanna[1], Alexandros G. Dimakis[1],
and Sahand Negahban[2]

[1]Department of Electrical and Computer Engineering
The University of Texas at Austin
{elenberg, rajivak}@utexas.edu, dimakis@austin.utexas.edu
[2]Department of Statistics
Yale Univeristy
sahand.negahban@yale.edu

June 19, 2016

Abstract

We connect high-dimensional subset selection and submodular maximization. Our results extend the work of Das and Kempe (2011) from the setting of linear regression to arbitrary objective functions. This connection allows us to obtain strong multiplicative performance bounds on several greedy feature selection methods without statistical modeling assumptions. This is in contrast to prior work that requires data generating models to obtain theoretical guarantees. 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.

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 $\boldsymbol{\beta}$ and consider problems of the form:

$$\bar{\boldsymbol{\beta}}^k \in \underset{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \leq k}{\arg\max}\, l(\boldsymbol{\beta}) \tag{1}$$

for some function l. This is a very general framework: the function l can be a linear regression R^2 objective, a generalized linear model (GLM) likelihood, a graphical model learning objective, or an arbitrary M-estimator [1]. This *subset selection* problem is NP-hard [2] even for the sparse linear regression objective, and a vast body of literature has analyzed different approximation algorithms under various assumptions.

The Restricted Isometry Property (RIP) and the (closely related) Restricted Eigenvalue (RE) property, are conditions on $l(\boldsymbol{\beta})$ that allow convex relaxations and greedy algorithms to solve the

subset selection problem within provable approximation guarantees. In parallel work, a several authors have demonstrated that the subset selection problem can be connected to submodular optimization [3, 4, 5, 6] 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 [7] for linear regression. Specifically, they showed that when $l(\boldsymbol{\beta})$ is the R^2 objective, it satisfies a weak form of submodularity when the linear measurements satisfy RIP. Note that for a given support of features S, the function $l(\boldsymbol{\beta}_{\mathsf{S}})$ 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 [6] in the linear regression setting discusses the notion of *suppressors*; however, that condition is stronger than the weak submodularity assumption.

In this paper we extend this machinery beyond linear regression, to any function $l(\boldsymbol{\beta})$. 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. [1] 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 [1] must be weakly submodular.

Using this result we establish strong multiplicative bounds on the performance of greedy algorithms, including (generalized) Orthogonal Matching Pursuit and Forward Stepwise Regression for general objective functions. These bounds allow us to provide sharp approximation bounds for 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 bounds from the best-case sparse solution. Furthermore, our results provide bounds for exact sparse recovery without any assumptions on the underlying model. Previous work on greedy methods provide guarantees under specific assumptions on the model. Convex methods such as ℓ_1 regularized objectives require extremely strong assumptions on the model, such as the irrepresentability conditions on the feature vectors, in order to provide exact sparsity guarantees.

Our main results are summarized below, with M, m, and γ defined formally in Section 2. If a function l has M-restricted smoothness (RSM) and m-restricted strong convexity (RSC), then the set function $f(\mathsf{S}) = -l(\boldsymbol{\beta}_{\mathsf{S}})$ is weakly submodular with parameter $\gamma \geq (m/M)^2$. We use this result to analyze three greedy algorithms, each progressively better but more computationally intensive: the Oblivious 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 keeping the feature that best improves the objective function at each step. We obtain the following performance bounds:

- After k steps, the Oblivious algorithm produces a (γ/k)-approximation to the best k-subset.
- After k steps, Orthogonal Matching Pursuit produces a $(1 - e^{-(m/4M)\gamma})$-approximation to the best k-subset.
- After k steps, Forward Stepwise Regression produces a $(1 - e^{-\gamma})$-approximation to the best k-subset.

Finally we 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. 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 [7] 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.

1.1 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, OMP, and CoSaMP [8]), forward-backward methods [9, 10], and truncated gradient methods [11]. Under the restricted strong convexity and smoothness assumptions that will be outlined below, 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 they relate 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 coefficient of the optimal coefficients as well as an upper-bound on the gradient of the objective at the optimal set of coefficients. Focusing explicitly on OMP, most previous results require the strong RIP assumption, whereas we only require the weaker RSC and RSM assumptions. Looking at OMP, results under arbitrary model conditions as we consider require the stronger RIP assumption as highlighted in Corollary 2 of Zhang [12]. 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 $\log[\log(p)k/n]$. Nevertheless, we emphasize that our results apply under arbitrary assumptions on the noise and use only RSC and RSM assumptions.

In [7], 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 [8] and statistical learning theory [13]. Greedy methods for sparsity constrained regression were analyzed in [9, 10, 11, 14, 15, 16] under assumptions similar to ours but without connections to submodularity. Convergence guarantees for ℓ_1 regularized regression were given for exponential families in [17], and for general nonlinear functions in [18]. 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. Submodularity has been used in the context of convex optimization [6] and active learning [3, 5]. In the latter, the focus is on selecting predictive data points instead of features.

2 Preliminaries

First we collect some notation that will be used throughout the remainder of this paper. Sets are represented by sans script fonts e.g. A, B. Vectors are represented using lower case bold letters e.g. $\mathbf{x}, \mathbf{y}$, and matrices are represented using upper case bold letters e.g. $\mathbf{X}, \mathbf{Y}$. The i-th column of $\mathbf{X}$ is denoted $\mathbf{X}_i$. Non-bold face letters are used for scalars e.g. j, M, r and function names e.g. $f(\cdot)$. The transpose of a vector or a matrix is represented by $\top$ e.g. $\mathbf{X}^\top$. Define $[p] := \{1, 2, \ldots, p\}$. For simplicity, we assume a set function defined on a ground set of size p has domain $[p]$. For singleton sets, we write $f(j) := f(\{j\})$. Next, we define the submodularity ratio of a set function $f(.)$.

Definition 1 (Submodularity Ratio [7]). *Let* $\mathsf{S}, \mathsf{L} \subset [p]$ *be two disjoint sets, and* $f(\cdot) : [p] \to \mathbb{R}$. *The submodularity ratio of* L *with respect to* S *is given by*

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

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

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

It is straightforward to show that f is submodular if and only if $\gamma_{\mathsf{L},\mathsf{S}} \geq 1$ for all sets L and S. In our application, $\gamma_{\mathsf{L},\mathsf{S}} \leq 1$ 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 [1, 19].

Definition 2 (Restricted Strong Concavity, Restricted Smoothness). *A function* $l : \mathbb{R}^p \to \mathbb{R}$ *is said to be restricted strong concave with parameter* m_Ω *and restricted smooth with parameter* M_Ω *if for all* $\mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^p$,

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

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 ML estimation where* $l(\cdot)$ *is the log-likelihood function and* $-l(\cdot)$ *is the data fit loss.*

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, *i.e.* normal, Bernoulli, Dirichlet, negative binomial, etc. [20]. Another interpretation is in minimizing the average Bregman divergence between the response y_i and the mean parameter $\langle \mathbf{x}_i, \beta \rangle$. There has been a large body of literature studying this method's statistical properties. These include establishing sparsistency, parameter consistency, and prediction error [17, 21, 1]. We refer the reader to the standard literature for more details on GLMs and exponential families [20, 22].

2.2 Support Selection Algorithms

We study general M-estimators of the form (1) for some function l. Note that l will implicitly depend on our specific data set, but we hide that for ease of notation. One common choice of l is the log-likelihood of a parametric distribution. [7] considers the specific case of maximizing R^2 objective. 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, then we wish to bound

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

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

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

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

Let $\boldsymbol{\beta}^{(\mathsf{A})}$ denote the $\boldsymbol{\beta}$ maximizing $f(\mathsf{A})$, and let $\boldsymbol{\beta}^{(\mathsf{A})}_{\mathsf{B}}$ denote $\boldsymbol{\beta}^{(\mathsf{A})}$ 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. 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 S is already selected, choose the feature with largest marginal gain, *i.e.* select $\{j\}$ such that $\mathsf{S} \cup \{j\}$ has the most improvement over 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 [23].

Algorithm 1 Oblivious Support Selection

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

Algorithm 2 Forward Stepwise Selection

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

Algorithm 3 Orthogonal Matching Pursuit

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

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*, features are selected to maximize the inner product with this residual, as shown in line 5 of Algorithm 3. 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.

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, the submodularity ratio is also bounded away from 0, which allows approximation guarantees for Algorithms 1–3. We make use of proof techniques similar to those of [7] which obtained approximation guarantees for support selections for linear regression. While our results are applicable to general functions, in the Appendix we discuss a direct application of maximum likelihood estimation for sparse generalized linear models.

We assume a differentiable function $l : \mathbb{R}^p \to \mathbb{R}$. Recall that we can define the equivalent, monotone set function $f : [p] \to \mathbb{R}$ for selected support as $\bar{f}(\mathsf{S}) = \max_{\mathrm{supp}(\boldsymbol{\beta}) \subset \mathsf{S}} l(\boldsymbol{\beta})$. The approximation guarantees we provide are on the *normalized* set function defined as $f(\mathsf{S}) = \bar{f}(\mathsf{S}) - \bar{f}(\emptyset)$. 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 $\gamma_{\mathsf{U},k}$ in terms of its strong concavity and smoothness parameters (see Definitions 1–2). Proofs of lemmas and theorems omitted from this section can be found in the Appendix.

Theorem 1 (RSC/RSM Implies Weak Submodularity)**.** *Define $f(\mathsf{S})$ as in (4), with a function $l(.)$ that is M-smooth and m-strongly concave on all $(|\mathsf{U}|+k)$-sparse vectors. Then the submodularity ratio $\gamma_{\mathsf{U},k}$ is lower bounded by*

$$\gamma_{\mathsf{U},k} \geq \left(\frac{m}{M}\right)^2. \tag{5}$$

Remark 2. *Since $(m/M)^2 \leq 1$, 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.*

3.1 Oblivious Algorithm

Next we generalize the results of [7], starting with the following lemma:

Lemma 1. *Let $1 \leq k \leq n$.*

$$f([k]) \geq \max\left\{\frac{1}{k}, \left(\frac{m}{2M}\right)^2\right\} \sum_{j=1}^{k} f(j).$$

Proof. Since $f(.)$ is monotone, $f(j) \leq f([k])$ for $j = 1, \ldots, k$. Summing over all j and dividing by k yields the first part of the inequality. The rest of the proof requires repeated applications of Definition 2 and is deferred to the Appendix. □

Now we present our first performance guarantee for feature selection.

Theorem 2 (Oblivious Algorithm Guarantee)**.** *Define $f(\mathsf{S})$ as in (4), with a function $l(.)$ that is M-smooth and m-strongly concave on all k-sparse vectors. 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*

$$f^{OBL} \geq \max\left\{\frac{m^2}{kM^2}, \frac{m^4}{4M^4}\right\} f^{OPT}. \tag{6}$$

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

$$f^{OBL} = f(\mathsf{S}) \geq C\sum_{j\in\mathsf{S}} f(j) \geq C\sum_{j\in\mathsf{S}^*} f(j) \geq C\gamma_{\emptyset,k} f(\mathsf{S}^*) \geq C\left(\frac{m}{M}\right)^2 f(\mathsf{S}^*) = C\left(\frac{m}{M}\right)^2 f^{OPT}.$$

□

3.2 Forward Stepwise Algorithm

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

Theorem 3 (Forward Stepwise Algorithm Guarantee)**.** *Define $f(\mathsf{S})$ as in (4), with a function that is M-smooth and m-strongly concave on all $2k$-sparse vectors. Let S_k^G be the set selected by the FS algorithm and S^* be the optimal set of size k corresponding to values f^G and f^{OPT}. Then*

$$f^G \geq \left(1 - e^{-\gamma_{\mathsf{S}_k^G,k}}\right) f^{OPT} \geq \left(1 - e^{-(m/M)^2}\right) f^{OPT}.$$

Proof. Let $l(.)$ 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 $f(\mathsf{S}_i^G) - f(\mathsf{S}_{i-1}^G)$ with $A(0) = 0$. Denote the remainder set at iteration i as $\mathsf{S}_i^R = \mathsf{S}^* \backslash \mathsf{S}_i^G$, and define $B(i) = f(\mathsf{S}^*) - f(\mathsf{S}_i^G)$, 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:*

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

Proof. Let $\mathsf{S} = \mathsf{S}_i^G$ 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 $\mathsf{S}^* \backslash \mathsf{S}$. S^R is a subset of the candidate variables available to the greedy algorithm at iteration $i+1$. Using Definition 1 and the fact that $k \geq |\mathsf{S}^R|$,

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

where the last inequality follows from the fact that $\mathsf{S} \cup \mathsf{S}^R \supseteq \mathsf{S}^*$. □

The rest of the proof follows standard results from submodular optimization [23, 24] and is deferred to the Appendix. □

Remark 3. *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 $r \to \infty$, however, since $\gamma_{\mathsf{S}_r^G,k}$ is a decreasing function of r.*

Corollary 1. *Let f^{G+} 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 $\gamma = \gamma_{\mathsf{S}_r^G,k}$ be the submodularity ratio associated with the output of f^{G+} and k. Then*

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

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 $f(\mathsf{S})$ as in (4), with a function $l(\cdot)$ that is M'-smooth and m'-strongly concave on all k-sparse vectors. Let 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*

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

Remark 4. *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 [7]).*

3.3 Orthogonal Matching Pursuit

OMP 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-smooth and m-strongly concave on all $2k$-sparse vectors. Let S_k^P be the set of features selected by the OMP algorithm and S_k be the optimal feature set on k variables corresponding to values f^{OMP} and f^{OPT}. Then*

$$f^{OMP} \geq \left(1 - e^{-(m/4M)\gamma_{\mathsf{S}_k^P,k}}\right) f^{OPT} \geq \left(1 - e^{-m^3/4M^3}\right) f^{OPT}.$$

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 $\gamma = (m/4M)\gamma_{\mathsf{S}_r^P,k}$ be the submodularity ratio associated with the output of f^{P+} and k. Then*

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

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.

4 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.

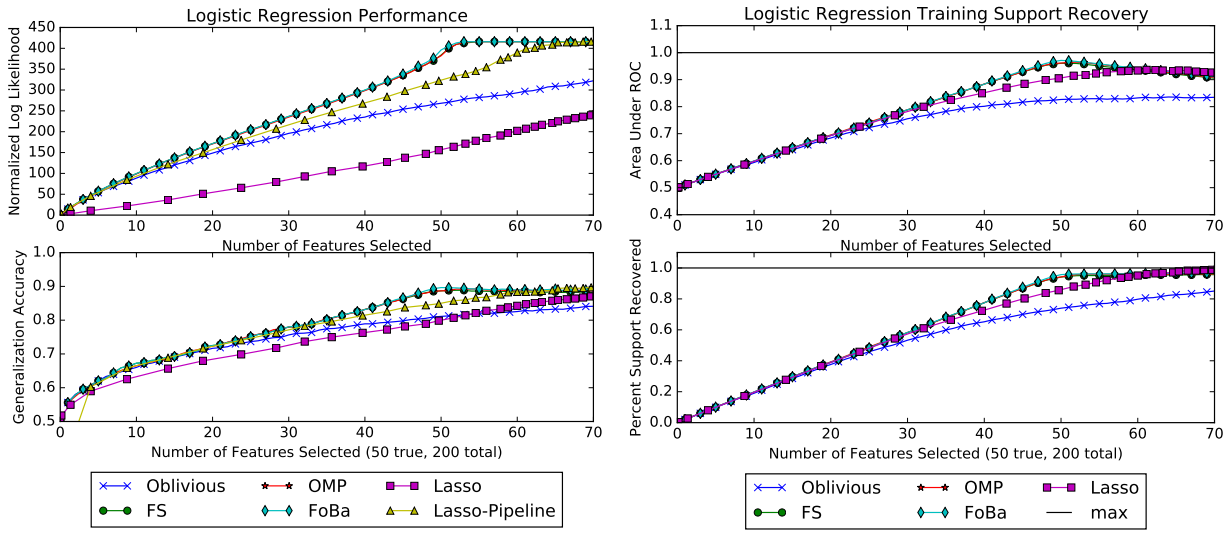
The Data: A synthetic experiment was conducted as follows: first each row of a 600×200 design matrix $\mathbf{X}$ is generated independently according to a first order AR process ($\alpha = 0.3$ and noise variance $\sigma^2 = 5$). This ensures that the features are heavily correlated with each other. 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 [25]. 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). 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) [26] 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 $s \in \{1, \ldots, 70\}$. 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.



(a) (b)

Figure 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.

Conclusions: We have extended the results of [7] 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

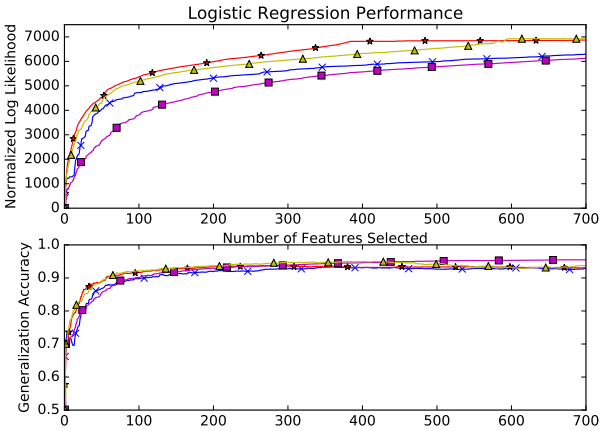


Figure 2: RCV1 Binary Dataset - $n = 10{,}000$, $p = 47{,}236$. OMP outperforms Lasso-Pipeline.

future work include similar analysis for other greedy algorithms that incorporate group sparsity or thresholding.

References

[1] S. Negahban, P. Ravikumar, B. Yu, and M. J. Wainwright, "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," *Statistica Sinica*, vol. 27, no. 4, 2012.

[2] B. K. Natarajan, "Sparse Approximate Solutions to Linear Systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[3] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch Mode Active Learning and its Application to Medical Image Classification," in *ICML*, 2006, pp. 417 – 424.

[4] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using Document Summarization Techniques for Speech Data Subset Selection," in *NAACL-HLT*, 2013, pp. 721–726.

[5] K. Wei, I. Rishabh, and J. Bilmes, "Submodularity in Data Subset Selection and Active Learning," *ICML*, pp. 1954–1963, 2015.

[6] F. R. Bach, "Learning with Submodular Functions: A Convex Optimization Perspective," *Foundations and Trends in Machine Learning*, vol. 6, 2013.

[7] A. Das and D. Kempe, "Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection," in *ICML*, 2011.

[8] D. Needell and J. A. Tropp, "CoSaMP : Iterative Signal Recovery from Incomplete and Inaccurate Samples," *Applied and Computational Harmonic Analysis*, vol. 3, no. 26, pp. 301–321, 2009.

[9] A. Jalali, C. Johnson, and P. Ravikumar, "On Learning Discrete Graphical Models Using Greedy Methods," in *NIPS*, 2011.

[10] J. Liu, J. Ye, and R. Fujimaki, "Forward-Backward Greedy Algorithms for General Convex Smooth Functions Over a Cardinality Constraint," in *ICML*, 2014, pp. 503–511.

[11] P. Jain, A. Tewari, and P. Kar, "On Iterative Hard Thresholding Methods for High-dimensional M-Estimation," in *NIPS*, 2014, pp. 685–693.

[12] T. Zhang, "Sparse Recovery With Orthogonal Matching Pursuit Under RIP," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6215–6221, September 2011.

[13] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore, "Approximation and Learning by Greedy Algorithms," *Annals of Statistics*, vol. 36, no. 1, pp. 64–94, 2008.

[14] A. C. Lozano, G. Świrszcz, and N. Abe, "Group Orthogonal Matching Pursuit for Logistic Regression," *Journal of Machine Learning Research*, vol. 15, pp. 452–460, 2011.

[15] A. Tewari, P. Ravikumar, and I. S. Dhillon, "Greedy Algorithms for Structurally Constrained High Dimensional Problems," in *NIPS*, vol. 24, 2011, pp. 1–10.

[16] X.-T. Yuan, P. Li, and T. Zhang, "Gradient Hard Thresholding Pursuit for Sparsity-Constrained Optimization," in *ICML*, 2014, pp. 1–26.

[17] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari, "Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity," in *AISTATS*, vol. 9, no. 5, 2010, pp. 381–388.

[18] Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang, "Sparse Nonlinear Regression: Parameter Estimation and Asymptotic Inference," *ICML*, pp. 1–32, 2016.

[19] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 559–616, Jan. 2015.

[20] L. D. Brown, "Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory," *Lecture Notes–Monograph Series*, vol. 9, 1986.

[21] S. A. van de Geer, "High-dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, vol. 36, pp. 614–645, 2008.

[22] D. K. Dey, S. K. Ghosh, and B. K. Mallick, Eds., *Generalized Linear Models: A Bayesian Perspective*, 1st ed., ser. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2000.

[23] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An An Analysis of Approximations for Maximizing Submodular Set Functions - I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[24] A. Krause and D. Golovin, "Submodular Function Maximization," *Tractability: Practical Approaches to Hard Problems*, vol. 3, pp. 71–104, 2014.

[25] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[26] T. Zhang, "Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models," *NIPS*, 2008.

[27] S. Bahmani, B. Raj, and P. T. Boufounos, "Greedy Sparsity-Constrained Optimization," *Journal of Machine Learning Research*, vol. 14, pp. 807–841, 2013.

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:

$$\begin{aligned}\mathbf{y} &= \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T \\ \mathbf{x}_1 &= \begin{bmatrix}0 & 1 & 0\end{bmatrix}^T \\ \mathbf{x}_2 &= \begin{bmatrix}z & \sqrt{1-z^2} & 0\end{bmatrix}^T \\ \mathbf{x}_3 &= \begin{bmatrix}2z & 0 & \sqrt{1-4z^2}\end{bmatrix}^T\end{aligned}$$

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,

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

which goes to zero as $z \to 0^+$. 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 $\{\mathbf{x}_1, \mathbf{x}_2\}$ ($S_2 = \{1, 2\}$).

B Additional Proofs

B.1 Proof of Theorem 1

Proof. We proceed by upper bounding the denominator and lower bounding the numerator of (2). The former follows from applying Defintion 2 with $\mathbf{x} = \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}$ and $\mathbf{y} = \boldsymbol{\beta}^{(\mathsf{L})}$, and then noting that $\nabla l(\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})})$ is zero on the coordinates of $\mathsf{L} \cup \mathsf{S}$:

$$l(\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \leq \frac{M}{2}\|\boldsymbol{\beta}^{(\mathsf{L})} - \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}\|_2^2. \tag{7}$$

Lower bounding the numerator requires more care. Similar to (7),

$$\frac{m}{2}\|\boldsymbol{\beta}^{(\mathsf{L})} - \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}\|_2^2 \leq l(\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}). \tag{8}$$

Next, we apply Definition 2 with $\mathbf{x} = \boldsymbol{\beta}^{(\mathsf{L})}$ and $\mathbf{y} = \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}$.

$$\frac{m}{2}\|\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2 \leq l(\boldsymbol{\beta}^{(\mathsf{L})}) - l(\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})}) + \langle\nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\rangle. \tag{9}$$

Comparing (9) and (8),

$$\langle\nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\rangle \geq m\|\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2. \tag{10}$$

Consider a single coordinate $j \in S$. The function at $\boldsymbol{\beta}^{(\mathsf{L}\cup\{j\})}$ is larger the function at any other $\boldsymbol{\beta}$ on the same support. In particular, $l(\boldsymbol{\beta}^{(\mathsf{L}\cup\{j\})}) \geq l(\mathbf{y}_j)$, where $\mathbf{y}_j := \boldsymbol{\beta}^{(\mathsf{L})} + \alpha\boldsymbol{\beta}_j^{(\mathsf{L}\cup\mathsf{S})}$ is $(|\mathsf{U}| + k)$-sparse. Applying Definition 2 with $\mathbf{x} = \boldsymbol{\beta}^{(\mathsf{L})}$ and $\mathbf{y} = \mathbf{y}_j$,

$$l(\boldsymbol{\beta}^{(\mathsf{L}\cup\{j\})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \geq l(\boldsymbol{\beta}^{(\mathsf{L})} + \alpha\boldsymbol{\beta}_j^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \geq \langle \nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \alpha\boldsymbol{\beta}_j^{(\mathsf{L}\cup\mathsf{S})}\rangle - \frac{M}{2}|\alpha\boldsymbol{\beta}_j^{(\mathsf{L}\cup\mathsf{S})}|^2$$

$$\Rightarrow \sum_{j\in S} l(\boldsymbol{\beta}_{\mathsf{L}\cup\{j\}}^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \geq \alpha\langle \nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \boldsymbol{\beta}_{\mathsf{S}}^{(\mathsf{L}\cup\mathsf{S})}\rangle - \frac{\alpha^2 M}{2}\|\boldsymbol{\beta}_{\mathsf{S}}^{(\mathsf{L}\cup\mathsf{S})}\|_2^2.$$

Next note that $\|\boldsymbol{\beta}_{\mathsf{S}}^{(\mathsf{L}\cup\mathsf{S})}\|_2^2 \leq \|\boldsymbol{\beta}_{\mathsf{S}}^{(\mathsf{L}\cup\mathsf{S})} + \boldsymbol{\beta}_L^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2$ and again $\langle \nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \boldsymbol{\beta}^{(\mathsf{L})}\rangle = 0$ since the supports are disjoint. This implies

$$\sum_{j\in S} l(\boldsymbol{\beta}_{\mathsf{L}\cup\{j\}}^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \geq \alpha\langle \nabla l(\boldsymbol{\beta}^{(\mathsf{L})}), \boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\rangle - \frac{\alpha^2 M}{2}\|\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2.$$

Combining with (10) and letting $\alpha = m/M$,

$$\sum_{j\in S} l(\boldsymbol{\beta}_{\mathsf{L}\cup\{j\}}^{(\mathsf{L}\cup\mathsf{S})}) - l(\boldsymbol{\beta}^{(\mathsf{L})}) \geq \left(\alpha m - \frac{\alpha^2 M}{2}\right)\|\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2 = \left(\frac{m^2}{2M}\right)\|\boldsymbol{\beta}^{(\mathsf{L}\cup\mathsf{S})} - \boldsymbol{\beta}^{(\mathsf{L})}\|_2^2. \tag{11}$$

Substituting (7) and (11) into (2), the result follows from taking the minimum over all sets L, S. □

B.2 Proof of Lemma 1

The rest of the proof requires applying Definition 2 to the underlying likelihood function l. Define a k-sparse $\overline{\boldsymbol{\beta}}$ by $\overline{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^{(j)}, j \in \mathsf{S}$ and 0 elsewhere. First, we apply Definition 2 with $\mathbf{x} = \mathbf{0}$ and $\mathbf{y} = \alpha\overline{\boldsymbol{\beta}}$ for some α. This implies

$$l(\alpha\overline{\boldsymbol{\beta}}) \geq \langle \nabla l(\mathbf{0}), \alpha\overline{\boldsymbol{\beta}}\rangle - \frac{M}{2}\|\alpha\overline{\boldsymbol{\beta}}\|_2^2. \tag{12}$$

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

$$\langle \nabla l(\mathbf{0}), \boldsymbol{\beta}^{(j)}\rangle \geq l(\boldsymbol{\beta}^{(j)}) + \frac{m}{2}|\boldsymbol{\beta}^{(j)}|^2 \geq \frac{m}{2}|\boldsymbol{\beta}^{(j)}|^2$$

$$\Rightarrow \langle \nabla l(\mathbf{0}), \overline{\boldsymbol{\beta}}\rangle \geq \frac{m}{2}\|\boldsymbol{\beta}^{(j)}\|_2^2. \tag{13}$$

Combining (12) and (13), and setting $\alpha = m/2M$,

$$l(\alpha\overline{\boldsymbol{\beta}}) \geq \alpha\frac{m}{2}\|\boldsymbol{\beta}^{(j)}\|_2^2 - \alpha^2\frac{M}{2}\|\overline{\boldsymbol{\beta}}\|_2^2 = \frac{m^2}{8M}\|\overline{\boldsymbol{\beta}}\|_2^2. \tag{14}$$

Applying Definition 2, with $\mathbf{x} = \boldsymbol{\beta}^{(j)}$ and $\mathbf{y} = \mathbf{0}$, noting that $\nabla l(\boldsymbol{\beta}^{(j)}) = 0$ on coordinate j, and again summing over $j \in \mathsf{S}$,

$$-l(\boldsymbol{\beta}^{(j)}) \geq -\frac{M}{2}|\boldsymbol{\beta}^{(j)}|^2 \quad \Rightarrow \quad \|\overline{\boldsymbol{\beta}}\|_2^2 \geq \frac{2}{M}\sum_{j\in\mathsf{S}} l(\boldsymbol{\beta}^{(j)}).$$

Combining this with (14), and noting that $l(\boldsymbol{\beta}^{(\mathsf{S})})$ optimizes l over all vectors with support in S,

$$f(\mathsf{S}) = l(\boldsymbol{\beta}^{(\mathsf{S})}) \geq l(\alpha\overline{\boldsymbol{\beta}}) \geq \left(\frac{m}{2M}\right)^2 \sum_{j\in\mathsf{S}} l(\boldsymbol{\beta}^{(j)}) = \left(\frac{m}{2M}\right)^2 \sum_{j=1}^{k} f(j).$$

This completes the proof.

B.3 Proof of Theorem 3

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 [23] or the survey [24]). Next, observe that $A(i+1) = B(i) - B(i+1)$. Combining this with Lemma 2 and letting $C = \gamma_{\mathsf{S}_k^G,k}/k$, 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 $B(k) = f^{OPT} - f^G$ and $B(0) = f^{OPT}$,

$$\begin{aligned} f^{OPT} - f^G &\leq (1-C)^k\, f^{OPT} \\ \Rightarrow f^G &\geq f^{OPT}\left[1-(1-C)^k\right] = f^{OPT}\left[1-\left(1-\frac{\gamma_{\mathsf{S}_k^G,k}}{k}\right)^k\right] \\ &\geq f^{OPT}\left(1-e^{-\gamma_{\mathsf{S}_k^G,k}}\right). \end{aligned}$$

The claim follows from applying Theorem 1.

B.4 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 strong concavity parameter at S_k. Let k' satisfy $M/m < k' < k$, where M is the maximum smoothness parameter of $l(.)$ at S_j taken over all $j \in [k', k]$. Then for large enough k,*

$$l(\boldsymbol{\beta}^{(\mathsf{S}_{k-1})}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)})\Theta\left(\left(\frac{k'}{k}\right)^{M/m}\right).$$

In particular with $M = \max_{\mathsf{S}} M(\mathsf{S})$, where $\mathsf{S} = \{\mathsf{S}_j \mid k/2 \leq j \leq k\}$,

$$l(\boldsymbol{\beta}^{(\mathsf{S}_{k/2})}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)})\Theta\left(2^{-M/m}\right).$$

Proof. Let j be the index that minimizes $|\boldsymbol{\beta}_j^{(\mathsf{S}_k)}|^2$. By M-smoothness and the fact that the min is smaller than the average,

$$\begin{aligned} l(\boldsymbol{\beta}^{(\mathsf{S}_{k-1})}) &\geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)}_{\mathsf{S}_k\setminus\{j\}}) \\ &\geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)}) + \langle \nabla l(\boldsymbol{\beta}^{(\mathsf{S}_k)}), \boldsymbol{\beta}^{(\mathsf{S}_k)}_{\mathsf{S}_k\setminus\{j\}} - \boldsymbol{\beta}^{(\mathsf{S}_k)}\rangle - \frac{M}{2}\|\boldsymbol{\beta}^{(\mathsf{S}_k)}_{\mathsf{S}_k\setminus\{j\}} - \boldsymbol{\beta}^{(\mathsf{S}_k)}\|_2^2 \\ &= l(\boldsymbol{\beta}^{(\mathsf{S}_k)}) - \frac{M}{2}|\boldsymbol{\beta}_j^{(\mathsf{S}_k)}|^2 \\ \Rightarrow \frac{l(\boldsymbol{\beta}^{(\mathsf{S}_{k-1})})}{l(\boldsymbol{\beta}^{(\mathsf{S}_k)})} &\geq 1 - \frac{M\|\boldsymbol{\beta}\|_2^2}{2k l(\boldsymbol{\beta}^{(\mathsf{S}_k)})} \end{aligned}$$

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

$$l(\boldsymbol{\beta}^{(\emptyset)}) - l(\boldsymbol{\beta}^{(\mathsf{S}_k)}) \leq -\frac{m}{2}\|\boldsymbol{\beta}^{(\mathsf{S}_k)} - \boldsymbol{\beta}^{(\emptyset)}\|_2^2 \Rightarrow -\frac{\|\boldsymbol{\beta}^{(\mathsf{S}_k)}\|_2^2}{l(\boldsymbol{\beta}^{(\mathsf{S}_k)})} \geq -\frac{2}{m}$$
$$\Rightarrow \frac{l(\boldsymbol{\beta}^{(\mathsf{S}_{k-1})})}{l(\boldsymbol{\beta}^{(\mathsf{S}_k)})} \geq 1 - \frac{M}{km}$$

Then applying iteratively for M/m constant, k large, and $M/m < k' < k$, as in [7] we have

$$l(\boldsymbol{\beta}^{(\mathsf{S}_{k-1})}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)}) \prod_{j=k'+1}^{k} 1 - \frac{M}{km} = l(\boldsymbol{\beta}^{(\mathsf{S}_k)})\Theta\left(\left(\frac{k'}{k}\right)^{M/m}\right).$$

□

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

$$l(\boldsymbol{\beta}^{(\mathsf{S}_k^G)}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_{k/2}^G)}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_{k/2})})\left(1 - e^{-\gamma_{\mathsf{S}_{k/2}^G, k/2}}\right)$$
$$\geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)})\Theta\left(2^{-M'/m'}\right)\left(1 - e^{-\gamma_{\mathsf{S}_{k/2}^G, k/2}}\right)$$
$$\Rightarrow l(\boldsymbol{\beta}^{(\mathsf{S}_k^G)}) \geq l(\boldsymbol{\beta}^{(\mathsf{S}_k)})\Theta\left(2^{-M'/m'}\right)\left(1 - e^{-(m'/M')^2}\right).$$

□

B.5 Proof of Theorem 5

Proof. The key idea at each step i is to relate the incremental gain from the index chosen by OMP to that of Greedy. Let $\mathsf{S} = \mathsf{S}_i^P$ be the set chosen by OMP up to iteration i. Given S, let v and u be the indices that would be selected by running one step of the OMP and Greedy algorithms, respectively. Define $D(i+1) = f(S_{i+1}^P) - f(S) = l(\boldsymbol{\beta}^{(\mathsf{S}\cup\{v\})}) - l(\boldsymbol{\beta}^{(\mathsf{S})})$, define $\tilde{A}(i) = f(S \cup u) - f(S) = l(\boldsymbol{\beta}^{(\mathsf{S}\cup\{u\})}) - l(\boldsymbol{\beta}^{(\mathsf{S})})$, and define $\tilde{B}(i) = f(\mathsf{S}^*) - f(\mathsf{S})$.

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 S^* by the following:*

$$D(i+1) \geq \frac{m}{4M}\frac{\gamma_{S_i^P, k}}{k}\tilde{B}(i).$$

Proof. We begin similar to the proof of Theorem 1. Let e_v be the unit vector with one at coordinate v. By Definition 2 with $\mathbf{x} = \boldsymbol{\beta}^{(S)}$ and $\mathbf{y} = \boldsymbol{\beta}^{(S)} + \eta e_v$ for any scalar η,

$$D(i+1) \geq l(\mathbf{y}) - l(\mathbf{x}) \geq \left\langle \nabla l(\boldsymbol{\beta}^{(S)}), \eta e_v \right\rangle - \frac{M}{2}|\eta|^2$$
$$\geq \left\langle \nabla l(\boldsymbol{\beta}^{(S)}), \eta e_u \right\rangle - \frac{M}{2}|\eta|^2,$$

since OMP chooses the coordinate which maximizes the gradient. In particular, let η be the u-th coordinate of $\alpha(\boldsymbol{\beta}^{(S\cup u)} - \boldsymbol{\beta}^{(S)})$, making ηe_u a scaling of the difference restricted to coordinate u. This implies

$$D(i+1) \geq \alpha\left\langle \nabla l(\boldsymbol{\beta}^{(S)}), (\boldsymbol{\beta}^{(S\cup u)} - \boldsymbol{\beta}^{(S)})_u \right\rangle - \frac{M}{2}|\eta|^2 = \alpha\left\langle \nabla l(\boldsymbol{\beta}^{(S)}), \boldsymbol{\beta}^{(S\cup u)} - \boldsymbol{\beta}^{(S)} \right\rangle - \frac{M}{2}|\eta|^2,$$

because the gradient is zero on S. Since $m \geq 0$, *i.e.* by concavity of $l(.)$,

$$\begin{aligned} D(i+1) &\geq \alpha \tilde{A}(i) + \alpha \frac{m}{2} \|\boldsymbol{\beta}^{(S \cup u)} - \boldsymbol{\beta}^{(S)}\|_2^2 - \frac{M}{2}|\eta|^2 \\ &\geq \alpha \tilde{A}(i) - \frac{M}{2}|\eta|^2. \end{aligned}$$

Next, from (7)

$$\begin{aligned} |\eta|^2 \leq \alpha^2 \|\boldsymbol{\beta}^{(S \cup u)} - \boldsymbol{\beta}^{(S)}\|_2^2 \leq \alpha^2 \frac{2}{m} \tilde{A}(i) \\ \Rightarrow D(i+1) \geq \left(\alpha \left(1 + \frac{m}{M}\right) - \alpha^2 \frac{M}{m}\right) \tilde{A}(i) \geq \left(\alpha - \alpha^2 \frac{M}{m}\right) \tilde{A}(i). \end{aligned}$$

Substituting $\alpha = \frac{m}{2M}$, and following the proof of Lemma 2 with $\mathsf{S}^R = \mathsf{S}^* \backslash \mathsf{S}_i^P$,

$$kD(i+1) \geq \frac{km}{4M} \tilde{A}(i) \geq \frac{m}{4M} \sum_{j \in \mathsf{S}^R} [f(\mathsf{S} \cup j) - f(\mathsf{S})] \geq \frac{m}{4M} \gamma_{\mathsf{S},k} \tilde{B}(i)$$

□

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

C 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 $\gamma_{\mathsf{U},k}$ 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 [22]:

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

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:

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

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

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

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

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

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

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

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

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 $\eta \mathbf{I} \preccurlyeq \mathbf{H}(\boldsymbol{\beta}) \preccurlyeq s\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I}$ for all $\boldsymbol{\beta}$.

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

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

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

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 1 (from [27])**.** *With Assumptions 1, for $\delta \in (0,1)$ and $n > \frac{R(\log r + r(1+\log \frac{p}{k} - \log \delta)}{\tilde{\theta}(1+\epsilon)\log(1+\epsilon) - \epsilon}$, $\lambda_{\max}(\mathcal{P}_\mathsf{T}^\top \mathbf{X}^\top \mathbf{X} \mathcal{P}_\mathsf{T}) \leq (1+\epsilon)\bar{\theta})$ with probability $(1-\delta)$.*

Corollary 3 (Regularized GLM Sample Complexity)**.** *Under the Assumption 1, for $\delta \in (0,1)$ and $n > \frac{R(\log r + r(1+\log \frac{p}{k} - \log \delta)}{\tilde{\theta}(1+\epsilon)\log(1+\epsilon) - \epsilon}$, with probability $(1-\delta)$, the submodularity ratio of f_{GLM1}, $\gamma_{\mathsf{U},r} \geq (\eta/(\eta + s(1+\epsilon)\bar{\theta}))^2$.*

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

Remark 5. *The above discussion can also be motivated by the Restricted Stability property of the Hessian of the loss function. Define:*

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

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

The Hessian is said to be μ_r-SRH (Stable Restricted Hessian) [27] if, $\mu_r \geq \frac{\mathbf{A}_r}{\mathbf{B}_r}$. It is straightforward to see that $\gamma_{\mathsf{U},r} \geq \frac{1}{\mu_r^2}$.

C.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 3. 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:

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

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

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

Restricted strong convexity for GLMs was studied by Negahban, et al. [1] 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. [19] extended the RSC conditions to hold uniformly for all k-sparse models for GLMs. We present the requisite results here.

Recall that $D_h(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$. 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 S if $\forall \mathbf{x}, \mathbf{y} \in \mathsf{S}$, $D(\mathbf{x}, \mathbf{y}) \geq \frac{m}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, 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 6 (from [19])**.** *If Assumptions 2 are 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 - $c_1 \exp(-c_2 n)$,*

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

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 6 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 4 (Sample complexity sub-gaussian design)**.** *Under Assumptions 2, for $n > \frac{c^2\sigma_{\mathbf{x}}}{\alpha_q^2}(r + |\mathsf{U}|)\log p$, the submodularity ratio for $f_{GLM2}(\cdot)$, $\gamma_{\mathsf{U},r} \geq (\frac{m'}{M'})^2$, where $m' = (\alpha_q - \frac{c^2\sigma_{\mathbf{x}}}{\alpha_q}\frac{(r+|\mathsf{U}|)\log p}{n}) > 0$, and $M' = \lambda_{\max}(\Sigma)(\frac{3}{2} + \frac{(r+|\mathsf{U}|)\log p}{n})$.*

Proof. Since $\|\Delta\|_1 \leq \sqrt{(r+|\mathsf{U}|)}\|\Delta\|_2$, from Theorem 6, $D(\mathbf{x},\mathbf{y}) \geq \frac{1}{2}(\alpha_q - \frac{c^2\sigma_{\mathbf{x}}}{\alpha_q}\frac{(r+|\mathsf{U}|)\log p}{n})\|\Delta\|_2^2$. The sample complexity bound follows by ensuring the RHS is > 0. This gives $h(\cdot)$ to be m-strongly convex with $m \geq (\alpha_q - \frac{c^2\sigma_{\mathbf{x}}}{\alpha_q}\frac{(r+|\mathsf{U}|)\log p}{n})$.

Similarly, a corresponding version of restricted smoothness by using the upper bound of $D(\mathbf{x},\mathbf{y})$ in Theorem 6 and using $\|\Delta\|_1 \leq \sqrt{r+|\mathsf{U}|}\|\Delta\|_2$. The expression for the submodularity ratio then follows from Theorem 1.

□