Restricted Strong Convexity Implies Weak Submodularity

Ethan R. Elenberg **Rajiv Khanna** **Alexandros G. Dimakis**
Department of Electrical and Computer Engineering
The University of Texas at Austin
{elenberg,rajivak}@utexas.edu dimakis@austin.utexas.edu

Sahand Negahban
Department of Statistics
Yale University
sahand.negahban@yale.edu

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. Our proof technique uses the concept of *weak submodularity* initially defined in [1]. We draw a connection between convex analysis and submodular set function theory which may be of independent interest, demonstrating that *Restricted Strong Convexity* (RSC) implies weak submodularity. More generally, this natural relaxation of submodularity can be used in other machine learning applications that have combinatorial structure.

For greedy feature selection, this connection allows us to obtain strong multiplicative performance bounds on several 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 $l(\cdot)$. 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 [2]. This *subset selection* problem is NP-hard [3] even for the sparse linear regression objective, and a vast body of literature has analyzed different approximation algorithms under various assumptions. In parallel work, several authors have demonstrated that the subset selection problem can be connected to submodular optimization [4–7] and that greedy algorithms are widely used for iteratively building good feature sets.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The Restricted Isometry Property (RIP) (and the closely related Restricted Eigenvalue property) is a condition on $l(\boldsymbol{\beta})$ that allows convex relaxations and greedy algorithms to solve the subset selection problem within provable approximation guarantees. The mathematical connection between submodularity and RIP was made explicit by Das and Kempe [1] 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 set of features S, the function $l(\boldsymbol{\beta}_{\mathsf{S}})$ 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.

In this paper we extend the machinery of [1] 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. [2] and is called Restricted Strong Convexity (RSC). The title of this paper should now be clear: we show that any objective function that satisfies RSC (and a natural smoothness assumption) 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 greedy algorithms in terms of only the objective function's strong convexity and smoothness. Our approach provides sharp approximation bounds in any setting where these fundamental structural properties are well-understood, *e.g.* 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 further over previous work by providing bounds on a solution that is guaranteed to match the desired sparsity, without any assumptions on the underlying 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 on the recovered solution. Our main result is summarized below, with M, m, and γ defined formally in Section 2.

Theorem 1 (RSC/RSM Implies Weak Submodularity, Informal)**.** *If a function $l(\cdot)$ has M-restricted smoothness (RSM) and m-restricted strong convexity (RSC), then the set function $f(\mathsf{S}) = \max_{\text{supp}(\boldsymbol{\beta}) \subseteq \mathsf{S}} -l(\boldsymbol{\beta})$ is weakly submodular with parameter $\gamma \geq {}^{m}/_{M}$.*

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 without accounting for dependencies or reduncancies. Orthogonal Matching Pursuit (OMP) greedily adds one feature at a time by picking the feature with the largest inner product with the function gradient at the current model. The gradient is the correct generalization of the residual error used in linear regression OMP. Finally, the most effective 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.

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 [1] 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.

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 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. Focusing explicitly on OMP, most previous results require the strong RIP assumption (such as in Corollary 2 of [12]), whereas we only require the weaker RSC and RSM assumptions. 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 $\log\left[\log(p)k/n\right]$.

Greedy algorithms are prevalent in compressed sensing literature [8], statistical learning theory [13], and sparsity constrained regression [9–11, 14–16] but without connections to submodularity. Submodularity has been used in the context of convex optimization [7] and active learning [4, 6]. In the latter, the focus is on selecting predictive data points instead of features. Recently, [17] and [18] proved constant factor guarantees for greedy algorithms using techniques from submodularity even though the problems considered were not strictly submodular.

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 $f(\cdot)$.

Definition 1 (Submodularity Ratio [1])**.** *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*

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

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 on performance of greedy selections.

Next we define the restricted versions of strong concavity and smoothness, consistent with [2, 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 $\mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^p$,*

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

If a function $l(\cdot)$ has restricted strong concavity parameter m, then its negative $-l(\cdot)$ has restricted strong convexity parameter m. In the case of ML estimation, for example, $l(\cdot)$ 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, let (m_k, M_k) denote the RSC and RSM parameters on the domain of all k-sparse vectors. If $j \leq k$, then $M_j \leq M_k$ and $m_j \geq m_k$. In addition, denote $\tilde{\Omega} := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x}-\mathbf{y}\|_0 \leq 1\}$ with corresponding smoothness parameter $\tilde{M}_1 \leq M_1$.

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. [1] 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, differentiable function $l : \mathbb{R}^p \to \mathbb{R}$, we can define an equivalent, monotone 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 consider the three support selection strategies for $f(\cdot)$ introduced in Section 1, which are widely used and simple to implement. Complete descriptions of the algorithms can be found in the full version of this paper [20].

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 $l(\cdot)$. Our approximation guarantees are on the *normalized* set function defined as $f(\mathsf{S}) = \bar{f}(\mathsf{S}) - \bar{f}(\emptyset)$. While our results are applicable to general functions, in [20] we discuss a direct application of maximum likelihood estimation for sparse generalized linear models. Theorem proofs and tighter guarantees are also deferred to [20].

Theorem 2 (RSC/RSM Implies Weak Submodularity, Formal)**.** *Define $f(\mathsf{S})$ as in* (4)*, with a function $l(\cdot)$ that is (m,M)-(strongly concave, smooth) on all $(|\mathsf{U}| + k)$-sparse vectors and $\tilde{M}_1$ smooth on all $(\mathbf{x}, \mathbf{y}) \in \tilde{\Omega}$. Then the submodularity ratio $\gamma_{\mathsf{U},k}$ is lower bounded by*

$$\gamma_{\mathsf{U},k} \geq \frac{m}{\tilde{M}_1} \geq \frac{m}{M}\,. \tag{5}$$

In the case of linear least-squares regression, m and M become sparse eigenvalues of the covariance matrix. Thus Theorem 2 is consistent with [1]. 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. Next we present performance guarantees for feature selection.

Theorem 3 (Oblivious Algorithm Guarantee)**.** *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 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*

$$f^{OBL} \geq \max\left\{\frac{m}{kM}, \frac{3m^2}{4M^2}, \frac{m^3}{M^3}\right\} f^{OPT}. \tag{6}$$

When the function is modular, *i.e.* $m = M$, then $f^{OBL} = f^{OPT}$ and the bound in Theorem 3 holds with equality. Next, we prove a stronger, guarantee for the greedy, Forward Stepwise algorithm.

Theorem 4 (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^G_k be the set selected by the FS algorithm and S^* be the optimal set of size k corresponding to values f^{FS} and f^{OPT}. Then*

$$f^{FS} \geq \left(1 - e^{-\gamma_{\mathsf{S}^G_k,k}}\right) f^{OPT} \geq \left(1 - e^{-m/M}\right) f^{OPT}. \tag{7}$$

This constant factor bound can be improved by running the Forward Stepwise algorithm for $r > k$ steps. The proof of Theorem 4 generalizes to compare performance of r greedy iterations to the optimal k-subset of features (see [20]). This generalized bound does not necessarily approach 1 as $r \to \infty$, however, since $\gamma_{\mathsf{S}^G_r,k}$ is a decreasing function of r.

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 4.

Theorem 5 (OMP Algorithm Guarantee)**.** *Define $f(S)$ as in* (4)*, with a log-likelihood function that is $\tilde{M}_1$-smooth on $\tilde{\Omega}$ and m-strongly concave on all $2k$-sparse vectors. Let f^{OMP} be the value of the set selected by OMP and f^{OPT} be the optimum value over all sets of size k. Then*

$$f^{OMP} \geq \left(1 - e^{-3m^2/4\tilde{M}_1^2}\right) f^{OPT}. \tag{8}$$

References

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

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

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

[4] 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.

[5] 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.

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

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

[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] J. Altschuler, A. Bhaskara, G. T. Fu, V. Mirrokni, A. Rostamizadeh, and M. Zadimoghaddam, "Greedy Column Subset Selection: New Bounds and Distributed Algorithms," in *ICML*, 2016.

[18] T. Horel and Y. Singer, "Maximization of Approximately Submodular Functions," in *NIPS*, 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] E. R. Elenberg, R. Khanna, A. G. Dimakis, and S. Negahban, "Restricted Strong Convexity Implies Weak Submodularity," https://arxiv.org/abs/1612.00804, 2016.