Greedy Minimization of Weakly Supermodular Set Functions

Edo Liberty[1] and Maxim Sviridenko[2]

1 **Yahoo Research**
sviri@yahoo-inc.com

2 **Amazon**
libertye@amazon.com

Abstract

This paper defines *weak-α-supermodularity* for set functions. It shows that minimizing such functions under cardinality constrains is a common task in machine learning and data mining. Moreover, any problem whose objective function exhibits this property benefits from a greedy extension phase. Explicitly, let S^* be the optimal set of cardinality k that minimizes f and let S_0 be an initial solution such that $f(S_0) \leq \rho f(S^*)$. Then, a greedy extension $S \supset S_0$ of size $|S| \leq |S_0| + \lceil \alpha k \ln(\rho/\varepsilon) \rceil$ yields $f(S) \leq (1+\varepsilon) f(S^*)$.

Example usages of this framework give streamlined proofs and new bi-criteria results for k-means, sparse regression, column subset selection, and sparse convex function minimization. Sparse regression and column subset selection are special cases of a new, more general, sparse multiple linear regression problem that is of independent interest. This paper also corrects a brittleness of the proof of Natarajan for the properties of the greedy algorithm for sparse regression.

1998 ACM Subject Classification G.1.3, G.1.6, G.4

Keywords and phrases Weak Supermodularity, Greedy Algorithms, Machine Learning, Data Mining

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Acces; Article No. 23; pp. 23:1–23:11

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Many problems in data mining and unsupervised machine learning take the form of minimizing a set function with cardinality constraints. More explicitly, denote by $[n]$ the set $\{1,\ldots,n\}$ and $f(S) : 2^{[n]} \to \mathbb{R}_+$. Our goal is to minimize $f(S)$ subject to $|S| \leq k$. These problems include clustering and covering problems as well as sparse regression, matrix approximation problems and many others. These combinatorial problems are hard to minimize in general. Finding good (e.g. constant factor) approximate solutions for them requires significant sophistication and highly specialized algorithms.

In this paper we analyze the behavior of the greedy algorithm to all of these problems. We start by claiming that the functions above are special. A trivial observation is that they are non-negative and non-increasing, that is, $f(S) \geq f(S \cup T) \geq 0$ for any $S, T \subseteq [n]$. This immediately shows that expanding solution sets is (at least potentially) beneficial in terms of reducing the function value. But, monotonicity is not sufficient to ensure that any number of greedy extensions of a given solution would significantly reduce the objective function.

To this end we need to somehow quantify the gain of adding a single element (greedily) to a solution set. Let $f(S) - f(S \cup T)$ be the reduction in f one gains by adding a set of elements T to the current solution S. Then, the average gain of adding elements from T *sequentially* is $[f(S) - f(S \cup T)]/|T \setminus S|$. One would hope that there exists an element in $i \in T \setminus S$ such $f(S) - f(S \cup \{i\}) \geq [f(S) - f(S \cup T)]/|T \setminus S|$. However, that would be false in general because different element contributions are not independent of each other. Nevertheless, it is true for supermodular functions (see Fact 2.1).

Combining this fact with the idea that T could be any set, including the optimal solution S^*, already gives some useful results for minimizing supermodular set functions. Specifically those for which $f(S^*)$ is bounded away from zero. Notice that k-means clustering (defined below) is exactly this kind of problem. Section 4 gives some new bicriteria results obtainable for k-means via the greedy extension algorithm of Section 3.

Alas, most problems of interest, such as regression, column subset selection, and feature selection are not supermodular. In Section 2 we define the notion of weak-α-supermodularity. Intuitively, weak-α-supermodular functions are those conducive to greedy type algorithms. The property requirers that there exists an element $i \in T \setminus S$ such that adding i *first* gains at least $[f(S) - f(S \cup T)]/\alpha|T \setminus S|$ for some $\alpha \geq 1$.

An analogous relaxation of the submodular property for set functions was considered in [6] (see definition 2.3). They define a *submodularity-ratio* for set functions which are not submodular. They show that if the *submodularity-ratio* is bonded, a greedy algorithm can be used to obtain bi-crateria results for the maximization problem. The work of [6] can be viewed as a direct extension of well know fact. Namely that the greedy algorithm provides a $(1 - 1/e)$-factor approximation for maximizing set functions $g(S)$ subject to $|S| \leq k$ if g for positive, monotone non-decreasing and submodular set functions [15].

This paper complements both [15] and [6] for the intuitively related process of greedily minimizing supermodular functions. While our setting is not significantly more complex it is quite different. In contrast to maximizing submodular functions, minimizing supermodular functions is, in general, hard [10]. The difficulty arrises from the fact that a value of zero of the objective function could force any constant factor approximation algorithm to find an optimal solution. Our work cannot overcome this fundamental (and unresolvable) difficulty.

We consider the case where either the objective function is bounded away from zero or one could obtain an approximate initial solution. In that case, supermodularity (or weak-α-supermodularity) is shown to be sufficient for obtaining good bi-crateria results using

the greedy algorithm. Section 2 includes notations and concepts that will be used throughout the paper. In section 3 we present two generic greedy algorithms and analyze their guaranties for weak-α-supermodular functions.

Many important problems in data mining and machine learning fall into this regime. As a warm-up, in Section 4 we obtain new bi-creteria results for k-means clustering, the objective function of which is supermodular. Section 5 presents the sparse multiple linear regression (SMLR) and shows that it is weakly-α-supermodular. We then streamline and slightly improve the result of [14] for sparse regression, also known as feature selection. Column Subset Selection (CSS) for matrix approximation is an instance of SMLR. Section 7 gives new bi-creteria results for CSS with little additional effort. Finally, we recreate the result of [16] for minimizing smooth and strongly convex functions with sparse solutions. The result is equivalent but the proof is simpler and shorter.

2 Preliminaries and definitions

Throughout the manuscript we denote by $[n]$ the set $\{1, \ldots, n\}$. We concern ourselves with non-negative set function $f(S) : 2^{[n]} \to \mathbb{R}_+$. More specifically monotone non-increasing set function such that $f(S) \geq f(S \cup T)$ for any two sets $S \subseteq [n]$ and $T \subseteq [n]$.

▶ **Definition 1.** A set function $f(S) : 2^{[n]} \to \mathbb{R}_+$ is said to be *supermodular* if for any two sets $S, T \subseteq [n]$

$$f(S \cap T) + f(S \cup T) \geq f(S) + f(T). \tag{1}$$

▶ **Definition 2.** A non-negative non-increasing set function $f(S) : 2^{[n]} \to \mathbb{R}_+$ is said to be *weakly-α-supermodular* if there exists $\alpha \geq 1$ such that for any two sets $S, T \subseteq [n]$

$$f(S) - f(S \cup T) \leq \alpha \sum_{i \in T \setminus S} \left(f(S) - f(S \cup \{i\}) \right). \tag{2}$$

This property is useful because we will later try to minimize f. It asserts that if adding $T \setminus S$ is beneficial then there is an element $i \in T \setminus S$ that contributes at least a fraction of that. The reason for the name of this property might also be explained by the following definition and lemma.

▶ Fact 2.1. A non-increasing non-negative supermodular function f is weakly-α-supermodular with parameter $\alpha = 1$.

Proof. For $S, T \subseteq [n]$ order the set $T \setminus S$ in an arbitrary order, i.e. $T \setminus S = \{i_1, \ldots, i_{|T \setminus S|}\}$. Define $R_0 = \emptyset$ and $R_t = \{i_1, \ldots i_t\}$ for $t > 0$. By supermodularity we have for any t

$$f(S) - f(S \cup \{i_t\}) \geq f(S \cup R_{t-1}) - f(S \cup R_{t-1} \cup \{i_t\}) \tag{3}$$

We note that $R_{t-1} \cup \{i_t\} = R_t$ and sum up Equation (3).

$$\begin{aligned} \sum_{t=1}^{|T \setminus S|} \left[f(S) - f(S \cup \{i_t\}) \right] &\geq \sum_{t=1}^{|T \setminus S|} f(S \cup R_{t-1}) - f(S \cup R_{t-1} \cup \{i_t\}) \\ &= f(S) - f(S \cup T) \ . \end{aligned}$$

Since $|T \setminus S| \cdot \max_{i \in T \setminus S} [f(S) - f(S \cup \{i\})] \geq \sum_{t=1}^{|T \setminus S|} [f(S) - f(S \cup \{i_t\})]$ this implies weak-1-supermodularily. ◀

3 General Greedy Extension Algorithms

We are given a weakly-α-supermodular set function $f(S)$ and would like to solve the following optimization problem

$$\min\{f(S) : |S| \leq k\}. \tag{4}$$

Let $0 < \Lambda_1 \leq \Lambda_2 \leq \dots$ be a non-decreasing bounded sequence of reals, i.e. $\max_t \Lambda_t < +\infty$. Our algorithm works in phases and we may assume that Λ_t is computed on step t of the algorithm. Consider a simple greedy algorithm that starts with some initial solution S_0 of value $f(S_0)$ (maybe $S_0 = \emptyset$) and sequentially and greedily adds elements to it to minimize f.

Algorithm 1 Greedy Extension Algorithm

input: Weakly-α-supermodular function $f(S)$, initial set S_0, parameters $k \in \mathbb{Z}_+$ and the sequence $\Lambda_1, \Lambda_2, \dots$
while $t \leq \lceil \alpha k \ln \Lambda_t \rceil$ **do**
  $S_t \leftarrow S_{t-1} \cup \arg\min_{i \in [n]} f(S_{t-1} \cup \{i\})$
output: S_t

Note that since the sequence Λ_t is bounded the algorithm terminates after at most $\lceil \alpha k \ln (\max_t \Lambda_t) \rceil$ iterations.

▶ **Lemma 3.** *Let S_τ be the output of Algorithm 1. Then $|S_\tau| \leq |S_0| + \lceil \alpha k \ln \Lambda_\tau \rceil$ and $f(S_\tau) \leq f(S^*) + \frac{f(S_0) - f(S^*)}{\Lambda_{\tau+1}}$ where S^* is an optimal solution of the optimization problem (4).*

Proof. The fact that $|S_\tau| \leq |S_0| + \lceil \alpha k \ln \Lambda_\tau \rceil$ is a trivial observation. For the second claim consider an arbitrary iteration $t \in [\tau]$ and consider the set $S^* \setminus S_{t-1}$. By monotonicity and weak α-supermodularity

$$\begin{aligned} f(S_{t-1}) - f(S^*) &\leq f(S_{t-1}) - f(S_{t-1} \cup S^*) \\ &\leq \alpha \cdot \sum_{i \in S^* \setminus S_{t-1}} f(S_{t-1}) - f(S_{t-1} \cup \{i\}) \\ &\leq \alpha k \cdot \max_{i \in [n]} f(S_{t-1}) - f(S_{t-1} \cup \{i\}) \\ &= \alpha k \cdot (f(S_{t-1}) - f(S_t)) \ . \end{aligned}$$

By rearranging the above equation and recursing over t we get

$$f(S_t) - f(S^*) \leq (f(S_{t-1}) - f(S^*))\,(1 - 1/\alpha k) \leq (f(S_0) - f(S^*))\,(1 - 1/\alpha k)^t$$

Substituting $\tau + 1 > \lceil \alpha k \ln \Lambda_{\tau+1} \rceil$ for the last step of the algorithm completes the proof.

$$\begin{aligned} f(S_\tau) - f(S^*) &\leq (f(S_0) - f(S^*))\,(1 - 1/\alpha k)^{\alpha k \ln \Lambda_{\tau+1}} \\ &\leq (f(S_0) - f(S^*))\, e^{-\ln \Lambda_{\tau+1}} \leq \frac{f(S_0) - f(S^*)}{\Lambda_{\tau+1}}. \end{aligned}$$

◀

▶ **Theorem 4.** *Let S_τ be the output of Algorithm 2 which is an instantiation of Algorithm 1 with parameters $\Lambda_t = f(S_0)/\varepsilon f(S_{t-1})$ for some error $\varepsilon \geq 0$. Then $|S_\tau| \leq |S_0| + \lceil \alpha k \ln(f(S_0)/\varepsilon f(S^*)) \rceil$ and $f(S_\tau) \leq f(S^*)/(1 - \varepsilon)$ where S^* is an optimal solution of the optimization problem (4).*

Algorithm 2 Greedy Extension Algorithm

input: Weakly-α-supermodular function $f(S)$, initial set S_0, $k \in \mathbb{Z}_+$
while $t \leq \lceil \alpha k \ln (f(S_0)/\varepsilon f(S_{t-1})) \rceil$ **do**
  $S_t \leftarrow S_{t-1} \cup \arg\min_{i \in [n]} f(S_{t-1} \cup \{i\})$
output: S_t

Proof. By Lemma 3 we have

$$|S_\tau| \leq |S_0| + \lceil \alpha k \ln \Lambda_\tau \rceil \leq |S_0| + \lceil \alpha k \ln(f(S_0)/\varepsilon f(S^*)) \rceil$$

and

$$\begin{aligned} f(S_\tau) &\leq f(S^*) + \frac{f(S_0) - f(S^*)}{\Lambda_{\tau+1}} \\ &= f(S^*) + \frac{f(S_0) - f(S^*)}{f(S_0)} \varepsilon f(S_\tau) \leq f(S^*) + \varepsilon f(S_\tau). \end{aligned}$$

◀

▶ **Theorem 5.** *Assume there exist a ρ-approximation algorithm creating S_0 such that $f(S_0) \leq \rho f(S^*)$. There exists an algorithm for generating S such that $|S| \leq |S_0| + \lceil \alpha k \left(\ln \frac{\rho}{\varepsilon}\right) \rceil$ and $f(S) \leq f(S^*)/(1-\varepsilon)$.*

Proof. Use the ρ-approximation algorithm to create S_0 for Algorithm 1 and apply Theorem 4. ◀

Algorithm 3 Greedy Extension Algorithm; an alternative stopping criterion

input: Weakly-α-supermodular function f, S_0, f_{stop}
repeat
  $S_t \leftarrow S_{t-1} \cup \arg\min_i f(S_{t-1} \cup \{i\})$
until $f(S_t) \leq f_{\text{stop}}$
output: $S = S_t$

▶ **Theorem 6.** *Let k' be the minimal cardinality of a set S' such that $f(S') \leq f'$. For any f_{stop} and an initial set S_0 such that $f' < f_{\text{stop}} < f(S_0)$ Algorithm 3 outputs S such that*

$$|S| \leq |S_0| + \left\lceil \alpha k' \left(\ln \frac{f(S_0) - f'}{f_{\text{stop}} - f'} \right) \right\rceil$$

Proof. Let $f' = f(S')$. The proof follows from Lemma 3 by setting $k = k_f$, $\Lambda_t = \frac{f(S_0)-f'}{f_{\text{stop}}-f'}$ and noticing that $\frac{f(S_0)-f'}{f_{\text{stop}}-f'} \leq \frac{f(S_0)-f}{f_{\text{stop}}-f}$. ◀

This alternative algorithm will be used in Section 6

4 k-means Clustering

As a gentle introduction we begin with deriving new bi-cretiria results for the k-means clustering problem. We begin by defining the constrained k means problem.

▶ **Definition 7** (Constrained k-means)**.** Given a set of n points $X \subset \mathbb{R}^d$, find a set $S \subset X$ minimizing $f(S) = \sum_{x \in X} \min_{x' \in S} \|x - x'\|^2$ subject to $|S| \leq k$.

▶ **Lemma 8.** *For the constrained k-means problem, one can find in $O(n^2dk\log(1/\varepsilon))$ time a set S of size $|S| = O(k) + k\log(1/\varepsilon)$ such that $f(S) \leq (1+\varepsilon)f(S^*)$ where $f(S^*)$ is the optimal solution.*

Proof. The constrained k-means objective function f is weakly-1-supermodular because it is supermodular (Fact 2.1). This is both well known and not hard to reverify. Using the algorithm of [2] one obtains a set S_0 of size $|S_0| = O(k)$ points from X for which $f(S_0) = O(f(S^*))$. Their technique improves on the analysis of well known k-means++ adaptive sampling scheme of [3]. Greedily extending S_0 and applying the analysis of Theorem 4 completes the proof. The quadratic dependency of the running time on the number of data points can be alleviated using the corset construction of [8, 9] ◀

▶ **Definition 9** (Unconstrained k-means)**.** Given a set of n points $X \subset \mathbb{R}^d$, find a set $S \subset \mathbb{R}^d$ minimizing $f(S) = \sum_{x\in X} \min_{c\in S} \|x-c\|^2$ subject to $|S| \leq k$.

▶ **Lemma 10.** *Let $f(S^*)$ be the optimal solution to the unconstrained k-means problem. One can find in time $O(n^2dk\log(1/\varepsilon))$ a set $S \in \mathbb{R}^d$ of size $|S| = O(k) + k\log(1/\varepsilon)$ such that $f(S) \leq (2+\varepsilon)f(S^*)$.*

Proof. The proof and the algorithm are identical to the above. The only point to note is that a $1+\varepsilon/2$ approximation to the constrained problem is at most a $2+\varepsilon$ approximation to the unconstrained one. See [3], for example, for the argument that the minimum of the constrained objective is at most twice that of the unconstrained one. ◀

Alternatively, we can utilize a more computationally expensive approach which goes through a reduction to the k-median problem.

▶ **Definition 11** (k-Median)**.** We are given a set X of data points, the set $\mathcal{C}$ of potential cluster center locations and the nonnegative costs $w_{ij} \geq 0$ for all $i,j \in X \times \mathcal{C}$. Find a set $S \subset \mathcal{C}$ minimizing $f(S) = \sum_{i\in X} \min_{j\in\mathcal{C}} w_{ij}$ subject to $|S| \leq k$.

It is known that given an instance (X,k) of the Unconstrained k-means problem one can construct in polynomial time an instance of the k-Median problem $(X,\mathcal{C},w,k)$ where $\mathcal{C} \subseteq \mathbb{R}^d$ such that for any solution of value Φ for the Unconstrained k-means problem there exists a solution of value $(1+\varepsilon)\Phi$ for the corresponding instance of the k-Median problem (see Theorem 7 [13]). Moreover, $|\mathcal{C}| = n^{O(\log(1/\varepsilon)/\varepsilon^2)}$. Therefore, after applying this transformation on our instance of the Unconstrained k-means and using the same initial solution S_0 as in Lemma 10 we derive.

▶ **Lemma 12.** *Let $f(S^*)$ be the optimal solution to the unconstrained k-means problem. One can find in time $O(n^{O(\log(1/\varepsilon)/\varepsilon^2)}dk)$ a set $S \in \mathbb{R}^d$ of size $|S| = O(k) + k\log(1/\varepsilon)$ such that $f(S) \leq (1+\varepsilon)f(S^*)$.*

5 Sparse Multiple Linear Regression

We begin by defining the Sparse Multiple Linear Regression (SMLR) problem. Given two matrices $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times\ell}$, and an integer k find a matrix $W \in \mathbb{R}^{n\times\ell}$ that minimizes $\|XW-Y\|_F^2$ subject to W having at most k non zero rows. We assume for notational brevity (and w.l.o.g.) that the columns of X have unit norm. An alternative and equivalent formulation of SMLR is as follows. Let X_S be a submatrix of the matrix X defined by the columns of X indexed by the set $S \subseteq [n]$. Let X_S^+ be the Moore-Penrose pseudo-inverse of X_S. It is well-known that the minimizer of $\|XW-Y\|_F^2$ subject to W

whose non zero rows are indexed by S is equal to $\|Y - X_S X_S^+ Y\|_F^2$. SMLR can therefore be reformulated as

$$\min_{S \subseteq [n]} \{ f(S) = \|Y - X_S X_S^+ Y\|_F^2 : |S| \leq k \} .$$

We can consequently apply our methodology from Section 3 to SMLR if we show that $f(S)$ is α-weakly-supermodular.

▶ **Lemma 13.** *For $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{m \times \ell}$ the SMLR minimization function $f(S) = \|Y - X_S X_S^+ Y\|_F^2$ is α-weakly-supermodular with $\alpha = \max_{S'} \|X_{S'}^+\|_2^2$.*

Proof. We first estimate $f(S) - f(S \cup T)$. Denote by $Z_{T \setminus S}$ the matrix whose columns are those of $X_{T \setminus S}$ projected away from the span of X_S and normalized. More formally, $\zeta_i = \|(I - X_S X_S^+) x_i\|$ and $z_i = (I - X_S X_S^+) x_i / \zeta_i$ for all $i \in T \setminus S$. Note that the column span of $Z_{T \setminus S}$ is orthogonal to that of X_S and that together they are equal to the column span of $X_{T \cup S}$. Using the Pythagorean theorem and the fact that $X_S X_S^+$ is a projection we obtain $f(S) = \|Y\|_F^2 - \|X_S X_S^+ Y\|_F^2$ and $f(S \cup T) = \|Y\|_F^2 - \|X_S X_S^+ Y\|_F^2 - \|Z_{S \setminus T} Z_{S \setminus T}^+ Y\|_F^2$. Substituting $T = \{i\}$ also gives $f(S) - f(S \cup \{i\}) = \|z_i z_i^T Y\|_F^2 = \|z_i^T Y\|_2^2$.

$$\begin{aligned}
f(S) - f(S \cup T) &= \|Z_{T \setminus S} Z_{T \setminus S}^+ Y\|_F^2 \\
&= \|(Z_{T \setminus S}^T)^+ \cdot Z_{T \setminus S}^T Y\|_F^2 \quad \text{By SVD} \\
&\leq \|(Z_{T \setminus S}^T)^+\|_2^2 \cdot \|Z_{T \setminus S}^T Y\|_F^2 \\
&= \|Z_{T \setminus S}^+\|_2^2 \cdot \sum_{i \in T \setminus S} \|z_i^T Y\|_2^2 \\
&\leq \|X_{T \cup S}^+\|_2^2 \cdot \sum_{i \in T \setminus S} \|z_i^T Y\|_2^2 \quad \text{See below} \\
&= \alpha \cdot \sum_{i \in T \setminus S} (f(S) - f(S \cup \{i\}))
\end{aligned}$$

For Equation (5) we use a non trivial transition, $\|Z_{T \setminus S}^+\|_2 \leq \|X_{T \cup S}^+\|_2$. By the definition of $Z_{T \setminus S}$ we can write for $i \in T \setminus S$ that $z_i = (x_i - \sum_{j \in S} \alpha_{ij} x_j)/\zeta_i$ and $\zeta_i = \|(I - X_S X_S^+) x_i\|$. For any vector w

$$Z_{T \setminus S} w = \sum_{i \in T \setminus S} x_i w_i / \zeta_i + \sum_{j \in S} x_j \sum_{i \in T \setminus S} w_i \alpha_{ij} / \zeta_i = X_{T \cup S} w'$$

where $w'_i = w_i / \zeta_i$ for $i \in T \setminus S$ and $w'_j = \sum_{i \in T \setminus S} w_i \alpha_{ij} / \zeta_i$ for $j \in S$. Since, $\zeta_i = \|(I - X_S X_S^+) x_i\| \leq \|x_i\| = 1$ we have $\|w'\| \geq \|w\|$. Finally, consider w such that $\|w\| = 1$ and $\|Z_{T \setminus S} w\| = \|Z_{T \setminus S}^+\|^{-1}$. This is the right singular vector corresponding to the smallest singular value of $Z_{T \setminus S}$. We obtain

$$\|Z_{T \setminus S}^+\|^{-1} = \|Z_{T \setminus S} w\| = \|X_{T \cup S} w'\| \geq \|X_{T \cup S}^+\|^{-1} \|w'\| \geq \|X_{T \cup S}^+\|^{-1} .$$

This completes the proof. ◀

▶ **Lemma 14.** *Let $f(S^*)$ be the optimal solution to the Sparse Multiple Linear Regression problem. One can find in time $O(\alpha k \log(\|Y\|_F^2/\varepsilon) \cdot n T_f)$ a set $S \subseteq [n]$ of size $|S| = \lceil \alpha k \log(\|Y\|_F^2/\varepsilon) \rceil$ such that $f(S) \leq f(S^*)/(1-\varepsilon)$ where T_f is the time needed to compute $f(S)$ once.*

6 Sparse Regression

The problem of Sparse Regression defined in [14] is an instance of SMLR where the number of columns in W and Y is $\ell = 1$. Since both W and Y are vectors we reduce to the more familiar form of this problem; minimize $\|Xw - y\|_2^2$ subject to $\|w\|_0 \leq k$.

Natarajan [14] analyzes the greedy algorithm for the sparse regression problem. He sets a desired threshold error E and defines k to be the minimum cardinality of a solution S^* that achieves $f(S^*) \leq E' = E/4$. He shows that for $\alpha = \max_{S'} \|X_{S'}^+\|^2$ the greedy algorithm finds a solution S such that $f(S) \leq E$ such that

$$|S| \leq \left\lceil 9k\alpha \ln \frac{\|y\|_2^2}{E} \right\rceil$$

6.0.0.1 Natarajan's implicit assumption:

In [14] Natarajan uses $\alpha = \|X^+\|^2$ instead of $\alpha = \max_{S'} \|X_{S'}^+\|^2$. This is only correct if the columns of X are linearly independent which seems to be an implicit assumption. In this setting $\alpha = \max_{S'} \|X_{S'}^+\|^2 = \|X^+\|^2$ by Cauchy's interlacing theorem. Note that $\max_{S'} \|X_{S'}^+\| \geq \|X^+\|$ if the columns of X are linearly dependent. This is the setting in the hardness result of [10] and is inevitable in the under constrained case where the number of columns is larger than their dimension.

Here, we apply Theorem 6 with initial solution $S_0 = \emptyset$ (which gives $f(S_0) = \|y\|_2^2$) and $E' = E/4$. It immediately yields that the greedy algorithm finds a solution of value $f(S) \leq E$ and

$$|S| \leq \left\lceil k\alpha \ln \frac{\|y\|_2^2 - E/4}{E - E/4} \right\rceil \leq \left\lceil \frac{4}{3} k\alpha \ln \frac{\|y\|_2^2}{E} \right\rceil .$$

using the inequality $\ln(\frac{4}{3}x - \frac{1}{3}) \leq \frac{4}{3} \ln x$ for $x \geq 1$. This improves the result of [14] in three ways

1. the approximation constant is smaller
2. its proof is more streamlined and
3. it extends to viability of the greedy algorithm to the under constrained case where the result of [14] does not hold.

7 Column Subset Selection Problem

Given a matrix X, Column Subset Selection (CSS) is concerned with finding a small set of columns whose span captures as much of the Frobenius norm of X. It was throughly investigated in the context of numerical linear algebra [5, 11, 12]. CSS can be formulated as follows, find a subset $S \in [n]$, $|S| \leq k$ of matrix columns that minimize $f(S) = \|X - X_S X_S^+ X\|_F^2$. This formulation makes it clear that this is a special case of SMLR where $Y = X$.

The work of [17] investigates the notion of a curvature $c \in [0, 1]$ for a nonincreasing set functions. They define it as follows:

$$c = 1 - \min_{j \in [n]} \min_{S,T \subseteq [n] \setminus \{j\}} \frac{f(S) - f(S \cup \{j\})}{f(T) - f(T \cup \{j\})}. \tag{5}$$

They show that there exists a greedy type algorithm that finds a solution of value at most $1/(1 - c)$ times the optimal value of the minimization problem for any objective set function with curvature c (Corollary 8.5 in [17]).

▶ **Lemma 15** (Lemma 9.1 from [17]). *Let $f(S)$ be the objective function for the Column Subset Selection Problem corresponding to the matrix X. The curvature c of $f(S)$ is such that $\frac{1}{1-c} \leq \kappa^2(X)$ where $\kappa(X)$ is the condition number of X.*

Note that for any matrix $X$with full column rank if $\tilde{X}$ is the matrix with normalized columns then $\|\tilde{X}^+\| \leq \kappa(X)$. We can find our initial solution S_0 by one of the three known methods:

1. an approximation algorithm from [17] finds a solution S_0 such that $|S_0| = k$ and performance guarantee $\rho = \kappa^2(X)$;
2. an approximation algorithm from [1, 7] with $|S_0| = k$ and $\rho = k + 1$;
3. an approximation algorithm from [4] with $|S_0| = 2k$ and $\rho = 2$;

▶ **Lemma 16.** *For the column subset selection problem for a column normalized matrix X and $\alpha = \max_{S'} \|X^+_{S'}\|_2^2$ one can find a set S such that*

$$f(S) \leq (1+\varepsilon) f(S^*) \qquad \text{and} \qquad |S| = O(k) + \alpha k \left(\ln \frac{\rho}{\varepsilon}\right).$$

Proof. Combining one of the above results with the algorithm from Section 3 completes the proof. ◀

8 Sparse Convex Function Minimization

One popular extension of the regression problem is to consider

$$f(S) = \min\{R(y) : \text{supp}(y) \subseteq S\} \tag{6}$$

where $R(y)$ is a convex function and $\text{supp}(y) = \{i \mid y_i \neq 0\}$. Following Shalev-Shwartz et al. [16], we consider a special case when the convex function $R(y)$ satisfies two additional conditions.

▶ **Definition 17.** A function $R(w)$ is said to be λ-strongly convex for $\lambda \geq 0$ if for each $w, u \in \mathbb{R}^d$ we have

$$R(w) \geq R(u) + \langle \nabla R(u), w - u \rangle + \frac{\lambda}{2} ||w - u||_2^2.$$

▶ **Definition 18.** A function $R(w)$ is said to be β-smooth if for each $w, u \in \mathbb{R}^d$ we have

$$R(w) \leq R(u) + \langle \nabla R(u), w - u \rangle + \frac{\beta}{2} ||w - u||_1^2.$$

Shalev-Shwartz et al. [16] gave many examples of such convex functions. In particular, they relate our Definition 18 to a class of functions arising in Machine Learning with β-smooth loss functions (see Lemma B1 and Section 3 in [16]).

▶ **Theorem 19.** *Given the set function $f(S)$ defined in (6) corresponding to β-smooth λ-strongly convex function $R(w)$. The set function $f(S)$ is α-weakly-supermodular with $\alpha = \frac{\beta}{\lambda}$.*

Proof. Let $y_S \in \mathbb{R}^d$ be a vector minimizing the function $R(y)$ among vectors with support S and $y_{S \cup T} \in \mathbb{R}^d$ be a vector minimizing function $R(y)$ among vectors with support $S \cup T$. For any vector $x \in \mathbb{R}^d$, let $x(j) \in \mathbb{R}$ be its j-th coordinate.

For each $j \in T \setminus S$, we define vector $\tilde{y}^j \in \mathbb{R}^d$ such that $\tilde{y}^j(j) = y_{S \cup T}(j)$ and $\tilde{y}^j(i) = 0$ for all $i \neq j$. It is enough to prove the inequality

$$R(y_S) - R(y_{S \cup T}) \leq \frac{\beta}{\lambda} \sum_{j \in T \setminus S} R(y_S) - R\left(y_S + \frac{\lambda}{\beta} \tilde{y}^j\right) \tag{7}$$

to prove the statement of the theorem. By applying Definitions 17 and 18 we derive

$$\begin{aligned}
\sum_{j\in T\setminus S} R(y_S) - R\left(y_S + \frac{\lambda}{\beta}\tilde{y}^j\right) &\geq \sum_{j\in T\setminus S}\left(-\left\langle \nabla R(y_S), \frac{\lambda}{\beta}\tilde{y}^j\right\rangle - \frac{\beta}{2}||\frac{\lambda}{\beta}\tilde{y}^j||_1^2\right) \\
&\geq -\frac{\lambda}{\beta}\left(\sum_{j\in T\setminus S}\left\langle \nabla R(y_S), \tilde{y}^j\right\rangle\right) - \frac{\lambda^2}{2\beta}||y_{S\cup T} - y_S||_2^2 \\
&= -\frac{\lambda}{\beta}\left\langle \nabla R(y_S), y_{S\cup T} - y_S\right\rangle - \frac{\lambda^2}{2\beta}||y_{S\cup T} - y_S||_2^2 \\
&\geq \frac{\lambda}{\beta}\left(R(y_S) - R(y_{S\cup T}) + \frac{\lambda}{2}||y_{S\cup T} - y_S||_2^2\right) - \frac{\lambda^2}{2\beta}||y_{S\cup T} - y_S||_2^2 \\
&= \frac{\lambda}{\beta}\left(R(y_S) - R(y_{S\cup T})\right)
\end{aligned}$$

where the first equality follows from the fact that $\nabla R(y_S)(j) = 0$ for all $j \in S$. ◀

Let R^* be the target value for our convex function $R(y)$ and k_f be the minimal cardinality of a set S' such that $f(S') \leq R^*$ where $f(S)$ is defined by (6). Combining Theorem 6 and Theorem 19 we derive

▶ **Theorem 20.** *For any $\varepsilon > 0$, let $f_{\text{stop}} = R^* + \varepsilon$ then the Algorithm 3 outputs S such that*

$$|S| \leq \left\lceil \frac{\beta}{\lambda} k_f \left(\ln \frac{R(\emptyset) - R^*}{\varepsilon}\right)\right\rceil.$$

The above theorem is analogous to Theorem 2.8 in [16].

9 Acknowledgments

We would like to thanks Christos Boutsidis for his contributions to early drafts of this paper, Sergei Vassilvitskii and Dan Feldman for their guidance and Petros Drineas for pointing out that Natarajan's proof would potentially not carry through for column subset selection.

References

1 A.Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 42th Annual ACM Symposium on Theory of Computing (STOC)*, 2010.

2 Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, pages 15–28, 2009.

3 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA*, pages 1027–1035, 2007.

4 C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

5 T.F. Chan and P. C.Hansen. Some applications of the rank revealing qr factorization. *SIAM Journal on Scientific and Statistical Computing*, 13:727, 1992.

6 A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *In Proceedings of ICML*, pages 1057–1064, 2011.

7 Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2:225–247, 2006.

8 Dan Feldman, Amos Fiat, Micha Sharir, and Danny Segev. Bi-criteria linear-time approximations for generalized k-mean/median/center. In *Proceedings of the Twenty-third Annual Symposium on Computational Geometry*, SCG '07, pages 19–26, New York, NY, USA, 2007. ACM. URL: `http://doi.acm.org/10.1145/1247069.1247073`, `doi:10.1145/1247069.1247073`.

9 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 569–578, New York, NY, USA, 2011. ACM. URL: `http://doi.acm.org/10.1145/1993636.1993712`, `doi:10.1145/1993636.1993712`.

10 Dean P. Foster, Howard J. Karloff, and Justin Thaler. Variable selection is hard. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 696–709, 2015. URL: `http://jmlr.org/proceedings/papers/v40/Foster15.html`.

11 G. H. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7:206–216, 1965.

12 M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong efficient algorithms for computing a strong rank-revealing qr-factorization. *SIAM Journal on Scientific Computing*, 17(848–869), 1996.

13 K. Makarychev, Y. Makarychev, M. Sviridenko, and J. Ward. A bi-criteria approximation algorithm for k means. In *submission*, 2015.

14 B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, April 1995.

15 G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978. URL: `http://dx.doi.org/10.1007/BF01588971`, `doi:10.1007/BF01588971`.

16 S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

17 Maxim Sviridenko, Jan Vondrak, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *In Proceedings of SODA 2015*, pages 1134–1148, 2014.