Greedy Minimization of Weakly Supermodular Set Functions

Christos Boutsidis* Edo Liberty† Maxim Sviridenko‡

Abstract

This paper defines *weak-α-supermodularity* for set functions. Many optimization objectives in machine learning and data mining seek to minimize such functions under cardinality constrains. We prove that such problems benefit 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 $f(S_0)/f(S^*) \leq \rho$. Then, a greedy extension $S \supset S_0$ of size $|S| \leq |S_0| + \lceil \alpha k \ln(\rho/\varepsilon) \rceil$ yields $f(S)/f(S^*) \leq 1+\varepsilon$. As example usages of this framework we give new bicriteria results for k-means, sparse regression, and columns subset selection.

1 Introduction

Many problems in data mining and unsupervised machine learning take the form of minimizing set functions 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 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 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, $f(S \cup T) \leq f(S)$ 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 enough to ensure that any number of greedy extensions of a given solution would 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 $f(S) - f(S \cup \{i\}) \geq [f(S) - f(S \cup T)]/|T \setminus S|$ but that would be false, in general, since the different element contributions are not independent of each other. Lemma 1, however, shows that this is true for supermodular functions.

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

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

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 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. A similar intuition gives a very famous result 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 [1].

*boutsidis@yahoo-inc.com, Yahoo Labs, New York, NY
†edo@yahoo-inc.com, Yahoo Labs, New York, NY
‡sviri@yahoo-inc.com, Yahoo Labs, New York, NY

Alas, most problems of interest, such as regression, columns subset selection, feature selection, and outlier detection (and many others) 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. Or, alternatively, the inequality above holds up to some constant $\alpha > 1$. Weak-α-supermodularity 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 $\alpha \geq 1$.

As an example for this framework we show in Section 5 that Sparse Multiple Linear Regression (SMLR) is weak-α-supermodular. Using this fact we extend (and slightly improve) the result of [2] for Sparse Regression and obtain new bicriteria results for Columns Subset Selection.

2 Weakly Supermodular Set Functions

In this section, we define our notation and the notion of *w*eak-α-supermodularity. Throughout the manuscript we denote by $[n]$ the set $\{1, \dots, n\}$. 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 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 $S, T \subseteq [n]$*

$$f(S) - f(S \cup T) \leq \alpha \cdot |T \setminus S| \cdot \max_{i \in T \setminus S} [f(S) - f(S \cup \{i\})] \, . \tag{2}$$

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.

Lemma 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. $T \setminus S = \{i_1, \dots, i_{|T \setminus S|}\}$. Define $R_0 = \emptyset$ and $R_t = \{i_1, \dots i_t\}$ 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).

$$\sum_{t=1}^{|T \setminus S|} [f(S) - f(S \cup \{i_t\})] \geq \sum_{t=1}^{|T \setminus S|} f(S \cup R_{t-1}) - f(S \cup R_{t-1} \cup \{i_t\}) = f(S) - f(S \cup T) \, .$$

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-supermodularly. □

3 Greedy Extension Algorithm

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

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

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.

Theorem 1. *Let S_τ be the output of Algorithm 1. Then $|S_\tau| \leq |S_0| + \lceil \alpha k \ln(f(S_0)/E) \rceil$ and $f(S_\tau) \leq f(S^*) + E$ where S^* is an optimal solution of the optimization problem* (4).

Algorithm 1 Greedy Extension Algorithm

```
input: Weakly-$\alpha$-supermodular function $f, S_0, k, E$
for $t = 1, \ldots, \lceil \alpha k \ln(f(S_0)/E) \rceil$ do
    $S_t \leftarrow S_{t-1} \cup \arg\min_{i \in [n]} f(S_t \cup \{i\})$
output: $S_t$
```

Proof. The fact that $|S_\tau| \leq |S_0| + \lceil \alpha k \ln(f(S_0)/E) \rceil$ 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

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

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 $\tau = \lceil \alpha k \ln(f(S_0)/E) \rceil$ for the last step of the algorithm completes the proof.

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

□

Theorem 2. *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 $f(S) \leq (1 + \varepsilon) f(S^*)$.*

Proof. Use the ρ-approximation algorithm to create S_0 for Algorithm 1 and set $E = \varepsilon f(S_0)$. □

Algorithm 2 Greedy Extension Algorithm; an alternative stopping criterion

```
input: Weakly-$\alpha$-supermodular function $f, S_0, f_{\text{stop}}$
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 3. *Let k_f be the minimal cardinality of a set S' such that $f(S') \leq f$. For any f_{stop} such that $f < f_{\text{stop}}$ Algorithm 2 outputs S such that*

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

Proof. The proof follows from Theorem 1 by setting $k = k_f$ and $E = f_{\text{stop}} - f$. □

4 Clustering

We will use the following auxiliary problem.

Definition 3 (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 well known that the objective function $f(S)$ of the k-Median problem is supermodular and therefore weakly-1-supermodular by Lemma 1. Our first application is a constrained version of the k-means clustering problem.

Definition 4 (Constrained k-Means)**.** *Given a set of 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 2. *Given a set of n points X define S^* the optimal solution to the constrained k-means problem. Namely, S^* minimizes $f(S)$ subject to $|S| \leq k$. One can find in $O(n^2 dk \log(1/\varepsilon))$ time a set S of size $|S| = O(k) + k \log(1/\varepsilon)$ such that $f(S) \leq (1+\varepsilon)f(S^*)$.*

Proof. The constrained k-means objective function f is weakly-1-supermodular because the problem is a special case of the k-Median problem defined above. Using the the algorithm of [3] one obtains a set S_0 of size $|S_0| = O(k)$ points from the data for which $f(S_0) = O(f(S^*))$. Their technique improves on the analysis of adaptive sampling method of [4]. Greedily extending S_0 and applying the analysis of Theorem 1 completes the proof. The quadratic dependency of the running time on the number of data points can be alleviated using the corset construction of [5, 6] □

The classical k-means clustering problem is defined as follows.

Definition 5 (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 3. *Let $f(S^*)$ be the optimal solution to the unconstrained k-means problem. One can find in time $O(n^2 dk \log(1/\varepsilon))$ 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 [4], 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. 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 [7]). 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 3 we derive.

Lemma 4. *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 $S \subseteq \{1, \ldots, n\}$. Let X_S^+ be the Moore-Penrose pseudo-inverse of the matrix X_S. It is well-known (and easy to verify) 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 5. *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 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 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 $f(S) - f(S \cup \{i\}) = \|z_i z_i^T Y\|_F^2$.

$$\begin{aligned}
f(S) - f(S \cup T) &= \|Z_{T\setminus S} Z_{T\setminus S}^+ Y\|_F^2 && \qquad (5)\\
&= \|(Z_{T\setminus S}^T)^+ \cdot Z_{T\setminus S}^T Y\|_F^2 && \text{by Singular Value Decomposition} \qquad (6)\\
&\le \|(Z_{T\setminus S}^T)^+\|_2^2 \cdot \|Z_{T\setminus S}^T Y\|_F^2 && \qquad (7)\\
&= \|Z_{T\setminus S}^+\|_2^2 \cdot \sum_{i\in T\setminus S} \|z_i^T Y\|_2^2 && \qquad (8)\\
&\le \|X_{T\cup S}^+\|_2^2 \cdot |T \setminus S| \max_{i\in T\setminus S} \|z_i^T Y\|_2^2 && \text{see below} \qquad (9)\\
&\le \alpha \cdot |T \setminus S| \, [f(S) - f(S \cup \{i\})] && \qquad (10)
\end{aligned}$$

For Equation (9) we use a non trivial transition, $\|Z_{T\setminus S}^+\|_2 \le \|X_{T\cup S}^+\|_2$. 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 $w \in \mathbb{R}^{|T\setminus S|}$

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

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, $\zeta_i = \|(I - X_S X_S^+)x_i\| \le \|x_i\| = 1$ we have $\|w'\| \ge \|w\|$. 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

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

Which completes the proof. □

Lemma 6. *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 $f(S) \le (1 + \varepsilon) f(S^*)$ where T_f is the time needed to compute $f(S)$ once.*

6 Sparse Regression

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

[2] analyzed the greedy algorithm for the sparse regression problem. He sets a desired threshold error E and defined k to be the minimum cardinality of a solution S^* that achieves $f(S^*) \le E' = E/4$. He showed that the greedy algorithm finds a solution S such that $f(S) \le E$ such that

$$|S| \le \left\lceil 9k \cdot \|X^+\|_2^2 \ln \frac{\|y\|_2^2}{E} \right\rceil$$

In his work [2] implicitly assumes the over constrained setting where the number of columns m in X is smaller than their dimension n and that X is full rank. In this setting $\alpha = \max_{S'} \|X_{S'}^+\| = \|X^+\|$ by Cauchy's interlacing theorem.

Here, we apply Theorem 3 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 $f(S) \le E$ such that

$$|S| \le \left\lceil k \cdot \|X^+\|_2^2 \ln \frac{\|y\|_2^2}{E - E/4} \right\rceil \le \left\lceil k \cdot \|X^+\|_2^2 \left(\ln \frac{\|y\|_2^2}{E} + \ln \frac{4}{3} \right) \right\rceil$$

This improves the result of [2] in three ways 1) the approximation factor is smaller by a constant factor 2) its proof is more streamlined and 3) it is extended to viability of the greedy algorithm to the under constrained case where the result of [2] does not hold. Specifically, where his implicit assumption that $\max_{S'} \|X^+_{S'}\| = \|X^+\|$ no longer holds.

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 [8, 9, 10]. In other words, find a subset $S \in [n]$, $|S| \leq k$ of matrix columns the 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 $Y = X$.

[11] investigated notion of a curvature $c \in [0, 1]$ for a nonincreasing set functions. They define it as follows:

$$c = 1 - \min_{j \in [n]} \min_{S,T \subseteq [n] \setminus \{j\}} \frac{f(S) - f(S \cup \{j\})}{f(T) - f(T \cup \{j\})}. \tag{11}$$

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 [11]).

Lemma 7 (Lemma 9.1 from [11])**.** *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 [11] finds a solution S_0 such that $|S_0| = k$ and performance guarantee $\rho = \kappa^2(X)$;
2. an approximation algorithm from [12, 13] with $|S_0| = k$ and $\rho = k + 1$;
3. an approximation algorithm from [14] with $|S_0| = 2k$ and $\rho = 2$;

Lemma 8. *For the columns subset selection problem for a column normalized matrix X and $\alpha = \max_{S'} \|X^+_{S'}\|_2^2$ one can fine a set S of value $f(S) \leq (1+\delta) f(S^*)$ such that*

$$|S| = O\left(\alpha k \left(\ln \frac{\rho}{\delta}\right)\right).$$

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

8 Acknowledgments

We would like to thanks Sergei Vassilvitskii and Dan Feldman for their guidance and for Petros Drineas and for pointing out the vulnerability of Natarajan's proof.

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] K. Makarychev, Y. Makarychev, M. Sviridenko, and J. Ward. A bi-criteria approximation algorithm for k means. In *submitted to COLT 2015.*, 2015.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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