ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413990

ON THE PERFORMANCE-COMPLEXITY TRADEOFF IN STOCHASTIC GREEDY WEAK SUBMODULAR OPTIMIZATION

Abolfazl Hashemi[‡], *Haris Vikalo*[†], *and Gustavo de Veciana*[†]

[†]Department of Electrical and Computer Engineering, UT Austin, TX, USA
[‡]Oden Institute for Computational Engineering and Sciences, UT Austin, TX, USA

ABSTRACT

Weak submodular optimization underpins many problems in signal processing and machine learning. For such problems, under a cardinality constraint, a simple greedy algorithm is guaranteed to find a solution with a value no worse than $1 - e^{-\gamma}$ of the optimal. Given the high cost of queries to large-scale signal processing models, the complexity of GREEDY becomes prohibitive in modern applications. In this work, we study the tradeoff between performance and complexity when one resorts to random sampling strategies to reduce the query complexity of GREEDY. Specifically, we quantify the effect of uniform sampling strategies on the performance through two criteria: (i) the probability of identifying an optimal subset, and (ii) the suboptimality of the solution's value with respect to the optimal. Building upon this insight, we propose a simple progressive stochastic greedy algorithm, study its approximation guarantees, and consider its applications to dimensionality reduction and feature selection tasks.

Index Terms— submodular optimization, greedy algorithm, subset selection, feature selection

1. INTRODUCTION

Technological advancements in various domains have enabled acquisition of high-dimensional datasets and have motivated vigorous research activities in the field of data sciences. High dimensionality of data presents computational and memory burdens and may adversely affect performance of the existing data analysis algorithms. Thus, it is desirable to arrive at a succinct yet information-preserving representation of the data. Submodular optimization [1] is a combinatorial optimization framework with desirable theoretical and practical properties; it has found applications in a number of settings including maximum weighted matching, facility location and coverage problems in discrete optimization [2], as well as active learning, influence maximization, and information gathering in machine learning [3–5]. In such problems, the goal is to maximize a monotonically increasing submodular function subject to a cardinality constraint characterizing the extent of representation, e.g., the number of features in a supervised learning task.

The objective function in some applications, e.g., sparse support selection and observation selection [6–9], is not necessarily a submodular function; rather, one deals with *weakly* submodular objectives that resemble diminishing return property of submodular functions. In many contemporary weak submodular maximization problems, one needs to handle increasingly larger quantities of data. The classical GREEDY algorithm for monotone γ-weak submodular maximization with cardinality constraint that enjoys an optimal $1 - e^{-\gamma}$ constant factor approximation [10] requires $\mathcal{O}(mk)$ function evaluations for cardinality constraint k and ground set of size m. Therefore, in data intensive applications where function evaluation is expensive, running GREEDY is infeasible. To this end, there have been recent efforts to exploit strong theoretical guarantees of GREEDY while improving on its complexity via resorting to either distributed and parallel computing schemes [11, 12], or methods to reduce the cost-per-iteration of GREEDY while remaining in centralized settings.

The focus of this paper is on the latter, i.e., weak submodular maximization centralized schemes. In particular, by relying on recent advances in design of uniform sampling strategies for weak submodular optimization [13, 14], we study the tradeoff between performance and complexity arising from reducing the cost-per-iteration of GREEDY via random restriction of the greedy search. Specifically, we quantify the effect of uniform sampling strategies on the performance of GREEDY through two criteria: (i) the probability of identifying an optimal subset, and (ii) the suboptimality of the solution's value with respect to the optimal. We show that although a fixed schedule of random restricted search spaces results in a nontrivial approximation factor with regard to the latter criterion, incremental increase of the size of the restricted greedy search is a necessary condition to identify the optimal subset in monotone weak submodular maximization tasks. This insight gives rise to a simple *progressive stochastic greedy* algorithm; we demonstrate its efficacy in dimensionality reduction and feature selection tasks.

2. WEAK SUBMODULAR MAXIMIZATION

A set function $f : 2^{\mathcal{X}} \to \mathbb{R}$ is monotone if $f(\mathcal{S}) \leq f(\mathcal{T})$ for all $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{X}$, where $\mathcal{X}$ denotes the so-called ground set. Furthermore, $f : 2^{\mathcal{X}} \to \mathbb{R}$ is submodular if

$$f(\mathcal{S} \cup \{j\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{j\}) - f(\mathcal{T}) \tag{1}$$

for all subsets $\mathcal{S} \subseteq \mathcal{T} \subset \mathcal{X}$ and $j \in \mathcal{X} \backslash \mathcal{T}$. The term $f_j(\mathcal{S}) = f(\mathcal{S} \cup \{j\}) - f(\mathcal{S})$ is the marginal value of adding element j to set $\mathcal{S}$.

Given a monotone non-decreasing set function $f : 2^{\mathcal{X}} \to \mathbb{R}$ with $f(\emptyset) = 0$, we are interested in solving the combinatorial optimization problem

$$\begin{aligned} \underset{\mathcal{S}}{\text{maximize}} \quad & f(\mathcal{S}) \\ \text{subject to} \quad & \mathcal{S} \subset \mathcal{X}, \ \ |\mathcal{S}| \leq k, \end{aligned} \tag{2}$$

which we denote by $\mathcal{P}(m, k)$, where $|\mathcal{X}| = m$. By a reduction to the well-known set cover problem, the combinatorial optimization (3) can be shown to be NP-hard [2, 15]. It has been shown in [10] that if $f(\cdot)$ is monotone and submodular, a simple greedy algorithm that iteratively selects an element with the highest marginal gain satisfies the optimal $1 - 1/e$ worst case approximation ratio.

In many problems, the objective function is not submodular but under certain conditions it behaves similarly. Such functions are called *weakly* submodular and the extent of their proximity to submodularity is captured by the submodularity ratio [8, 9]. The submodularity ratio of a normalized and monotone non-decreasing function f with respect to set $\mathcal{T}$ and parameter $k \geq 1$ is defined as

$$\gamma_{\mathcal{T},k} = \underset{\mathcal{L},\mathcal{S}\in\mathcal{X}}{\text{minimize}} \quad \frac{\sum_{j\in\mathcal{S}} f(\mathcal{L}\cup\{j\}) - f(\mathcal{L})}{f(\mathcal{L}\cup\mathcal{S}) - f(\mathcal{L})} \quad \text{subject to} \quad \mathcal{L}\subseteq\mathcal{T}, |\mathcal{S}|\leq k, \mathcal{S}\cap\mathcal{L}=\emptyset, \tag{3}$$

i.e., it captures how much more can f increase by adding any subset $\mathcal{S}$ of size k to $\mathcal{L}$ compared to the combined benefits of adding its individual elements to $\mathcal{L}$. Note that a set function f is submodular if and only if $\gamma_{\mathcal{T},k} \geq 1$. Formally, a set function f is weak submodular if $0 < \gamma_{\mathcal{T},k} < 1$. It is worth pointing out other weak submodularity notions, such as those in [16,17], which depending on the application may simplify the derivation of approximation bounds (see e.g., [18–20]).

By relying on the submodularity ratio, one can extend the theoretical results of [10] for GREEDY to the case of weak submodular functions [8], yielding

$$f(\mathcal{S}_g) \geq \left(1 - e^{-\gamma_{\mathcal{S}_g,k}}\right) f(\mathcal{S}^\star), \tag{4}$$

where $\mathcal{S}_g$ is the subset selected when solving (2) approximately via GREEDY, $\gamma_{\mathcal{S}_g,k}$ denotes the submodularity ratio defined in (3), and $\mathcal{S}^\star$ with $|\mathcal{S}^\star| = k$ denotes the optimal subset.

The approximation result (4) implies that if the objective function is monotone and weak submodular, the greedy selection scheme which in each iteration selects an element with the highest marginal gain finds a solution that is close to the optimal.

Finally, since we further aim to study conditions for the exact identification of $\mathcal{S}^\star$, we formally state the following definition.

Definition 1. *Let* ALG *be an approximation algorithm for the weak submodular optimization problem* (2) *with a unique solution* $\mathcal{S}^\star$. *Let* $\mathcal{S}_{alg}$ *be the output of* ALG. *Then,* ALG *successfully identifies* $\mathcal{S}^\star$ *if* $\mathcal{S}_{alg} = \mathcal{S}^\star$. *Furthermore, the probability of success of* ALG *is defined as* $\Pr(\mathcal{S}_{alg} = \mathcal{S}^\star)$.

3. PERFORMANCE-COMPLEXITY TRADEOFFS

Running GREEDY can be computationally expensive for large datasets. This is because if $|\mathcal{X}| = m$, in each of k iterations of GREEDY one needs to find the marginal gain of $\mathcal{O}(m)$ elements. Although computational costs can be reduced using the so-called lazy evaluations [1], the worst case number of function evaluations of GREEDY is $\mathcal{O}(mk)$. The prohibitive complexity of GREEDY for large-scale datasets has motivated the design of more efficient schemes for weak submodular maximization. A simple strategy to reduce the number of function evaluations is to restrict the search domain in each iteration of the greedy selection procedure using uniform random sampling (see Algorithm 1). Specifically, in the i^{th} iteration, instead of evaluating marginal gains of all elements, one restricts the evaluation to a smaller subset $\mathcal{R}^{(i)}$ of cardinality r_i. The simplest variants of such a strategy use a fixed schedule of the search space sizes $r_i = r$ in every iteration [13, 14, 19]. In this section, we aim to quantify the impact of r_i on the performance of Algorithm 1. To this end, we first establish a lower bound on the expected worst case performance of Algorithm 1 in terms of the suboptimality of the returned value (Theorem 1), and then study the probability of exact identification of the optimal subset (Theorem 2).[1]

Algorithm 1 GREEDY with restricted uniform search space

1: **Input:** Weak submodular function f, ground set $\mathcal{X}$, number of elements to be selected k, search space schedule $\{r_i\}_{i=0}^{k-1}$.
2: **Output:** Subset $\mathcal{S}^{(k)} \subseteq \mathcal{X}$ with $|\mathcal{S}^{(k)}| = k$.
3: Initialize $\mathcal{S}^{(0)} = \emptyset$
4: **for** $i = 0, \ldots, k-1$ **do**
5: Form $\mathcal{R}^{(i)}$ by sampling $\min(r_i, m)$ elements from $\mathcal{X}$ uniformly at random.
6: $j_s \in \text{argmax}_{j\in\mathcal{R}^{(i)}} f_j(\mathcal{S}^{(i)})$
7: $\mathcal{S}^{(i+1)} = \mathcal{S}^{(i)} \cup \{j_s\}$
8: **end for**

Theorem 1. *Let* $\mathcal{S}^{(k)}$ *denote the random subset selected by Algorithm 1 using schedule* $\{r_i\}$, *and let* $\gamma_{\mathcal{S}^{(k)},k}$ *be the submodularity ratio of the set function objective in* (2) *with respect to* $\mathcal{S}^{(k)}$. *Then*

$$\mathbf{E}[f(\mathcal{S}^{(k)})] \geq \left(1 - e^{-\gamma_{\mathcal{S}^{(k)},k}} - \gamma_{\mathcal{S}^{(k)},k} e^{-\frac{\bar{r}k\eta}{m}}\right) f(\mathcal{S}^\star), \tag{5}$$

where $\bar{r} = \min_i r_i$ *and* $\eta = 1 + \mathcal{O}(1/k)$.

Theorem 1 establishes that Algorithm 1, up to a negligible term $\gamma_{\mathcal{S}^{(k)},k} e^{-\frac{\bar{r}k\eta}{m}}$, enjoys an approximation factor that is nearly identical to that of GREEDY (see (4)); the latter requires evaluation of all elements in $\mathcal{X}$ in each iteration. To see this, assume the objective function in (2) is submodular (i.e., $\gamma_{\mathcal{S}^{(k)},k} = 1$) and $\eta \approx 1$. Then, we may approximately write

$$\mathbf{E}[f(\mathcal{S}^{(k)})] \geq \left(1 - \frac{1}{e} - e^{-\frac{\bar{r}k}{m}}\right) f(\mathcal{S}^\star), \tag{6}$$

where $e^{-\frac{\bar{r}k}{m}}$ is the cost of reducing the complexity via restricted uniform search spaces. Furthermore, note that a larger schedule $\{r_i\}$ evidently improves the approximation factor in (5). This analysis thus suggests the use of Algorithm 1 to reduce the cost of GREEDY.

At first glance, Theorem 1 may appear to suggest that any variant of Algorithm 1 with a restricted uniform search space could perform similarly to GREEDY. However, in Theorem 2 below we show that, somewhat surprisingly, for large-scale problems some variants with overwhelming probability fail to successfully identify the optimal subset.

Theorem 2. *Consider a sequence of optimization problems* $\mathcal{P}(m,k)$ *in* (2) *under an increasingly higher dimensional settings, i.e., the setting where* $m, k \to \infty$, $m > k$. *Let* ALG *denote a variant of* GREEDY *with a restricted uniform search space* $\mathcal{R} \subset [m]$ *having cardinality* r, *i.e.,* r *denotes the number of oracle calls in each iteration of* ALG. *The following claims hold:*

1. *If there exists* $\alpha \in (0,1)$ *such that* $r \leq k^{\alpha-1}m$, *then the probability that* ALG *succeeds on* $\mathcal{P}(m,k)$ *goes to zero, i.e.,*

$$\lim sup_{m,k\to\infty} \Pr\left(\mathcal{S}_{alg}^{(k)} = \mathcal{S}^\star\right) = 0. \tag{7}$$

2. *If there exists* $\alpha_1 \in (0,1)$ *such that* $r \leq \alpha_1 m$, *then the probability that* ALG *succeeds on* $\mathcal{P}(m,k)$ *satisfies*

$$\lim sup_{m,k\to\infty} \Pr\left(\mathcal{S}_{alg}^{(k)} = \mathcal{S}^\star\right) \in (\delta_1, \delta_2), \tag{8}$$

[1] The proofs are omitted for brevity and can be found in the extended version of the paper [21].

where δ_1 and δ_2 are positive constants that depend on α_1 such that $0 < \delta_1 < \delta_2 < 0.63$.[2]

Theorem 2 establishes upper bounds on the probability that a variant of GREEDY with a restricted search space constructed uniformly at random identifies $\mathcal{S}^\star$ exactly in two scenarios: (i) If the size of the search space remains fixed in each iteration of ALG and the algorithm makes $\mathcal{O}(mk^\alpha)$ oracle calls for some $\alpha \in (0,1)$, then the probability of the exact identification approaches zero as the problem dimension grows. (ii) If the size of the search space remains fixed in each iteration of ALG and strictly less than $[m]$, and the algorithm makes $\mathcal{O}(mk)$ oracle calls, then although the probability of the exact identification does not approach zero, it is not asymptotically one either. Note that in many applications, including sparse reconstruction and sparse learning [22,23], an arbitrarily high success probability is a condition required to establish any nontrivial sample complexity results, i.e. the minimum number of data points for successful recovery and prediction. Therefore, the two parts of Theorem 2 collectively imply that having an increasing *schedule* of search spaces which ultimately reaches m is a necessary condition to exactly identify the optimal support $\mathcal{S}^\star$ with high probability.

4. PROGRESSIVE STOCHASTIC GREEDY

Based on the insights of Theorem 2, we propose a simple variant of Algorithm 1, referred to as *progressive stochastic greedy* (PSG). PSG is designed following the idea that in order to identify the optimal support $\mathcal{S}^*$ exactly, the number of oracle calls in each iteration of Algorithm 1 need not be equal. Specifically, in the early iterations the search space can be drastically reduced to a small subset which with high probability contains at least one index from $\mathcal{S}^*$. However, in the subsequent iterations, assuming that the algorithm has been accurately identifying elements of $\mathcal{S}^*$, the search domain needs to be as large as $O(m)$ to allow the possibility of including an element from $\mathcal{S}^*$. That is, since the goal is to identify exactly all the elements of $\mathcal{S}^\star$, one should *progressively* increase the size of the search set thus improving the probability of success.

To this end, PSG employs an intuitive progression of the search set size. Specifically, in the i^{th} iteration the proposed scheme samples $r_i = \frac{m}{k-i}\log\frac{1}{\epsilon}$ elements uniformly at random from $[m]$ to construct the search set $\mathcal{R}^{(i)}_{psg}$. Here ϵ, selected such that $e^{-k} \leq \epsilon \leq e^{-\frac{k}{m}}$, is a parameter that allows one to strike a desired balance between the performance and complexity. It should be noted that in practice the sampling may be with or without replacement. Additionally, since it should hold that $r_i \leq m$ for all $i = 0, \ldots, k-1$, for any iteration i such that $i \geq k - \log\frac{1}{\epsilon}$ we set r_i to its maximum value, m.

Remark 1. *A schedule of search spaces with cardinality $r_i = \frac{m}{k-i}\log\frac{1}{\epsilon}$, explored by* PSG*, satisfies the necessary condition to exactly identify the optimal support $\mathcal{S}^\star$ with high probability and hence the result of Theorem 2 does not apply there.*

5. VERIFYING THE THEORY

In this section, we verify our theoretical results by comparing them to the empirical ones obtained via Monte Carlo (MC) simulations. Specifically, we consider the task of sparse support selection [22,23] where we are given a linear measurement model $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^m$ is a k-sparse unknown vector, i.e., a vector with at most k non-zero components, $\mathbf{y} \in \mathbb{R}^n$ denotes the vector of measurements, $\mathbf{A} \in \mathbb{R}^{n\times m}$ is the coefficient matrix assumed to be full rank, and $\boldsymbol{\nu} \in \mathcal{R}^n$ denotes the additive measurement noise vector. The search for a sparse approximation of $\mathbf{x}$ leads to the NP-hard cardinality-constrained least-squares problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \qquad \text{subject to} \quad \|\mathbf{x}\|_0 \leq k, \tag{9}$$

which can be interpreted as an instance of (2) [8, 19].

We consider a setting with increasing support size k (varied from 10 to 100) and set the dimension of the signal and the number of measurements to $m = 2k^{1.5}$ and $n = 6k\log(m/k\sqrt[6]{4\beta})$, respectively, for three different values of $\beta = 0.1, 0.05, 0.01$. In each trial, we select locations of the nonzero elements of $\mathbf{x}$ uniformly at random and draw those elements from a normal distribution. Entries of the coefficient matrix $\mathbf{A}$ are also generated randomly from $\mathcal{N}(0, \frac{1}{n})$. The results are averaged over 1000 MC trials. Note that, as we show in the extended version of the paper [21], in the above settings PSG is able to recover $\mathbf{x}$ exactly.

First, we investigate the exact performance of PSG with the schedule

$$r_i = \frac{m}{k-i}\log\frac{1}{\epsilon}, \quad \epsilon = \frac{\beta}{k} \tag{10}$$

for $\beta = 0.1, 0.05, 0.01$, and show the results in Fig. 1(a). As can see from the figure, the empirical exact recovery rate of PSG is very close to one; this coincides with the theoretical lower bound of $1 - 2\beta$ established in the extended manuscript [21] that builds upon the insights of Theorem 2 (i.e, the achieved rate is $0.8, 0.9, 0.98$ for $\beta = 0.1, 0.05, 0.01$, respectively).

Next, we empirically verify the results of Theorem 2 wherein we established an upper bound on the success probability of a variant of Algorithm 1, named ALG, with a restricted uniform search space. Fig. 1 compares this theoretical result with the empirical success rate for $r = m/\sqrt{k}$ and $r = m/2$, which correspond to instances of the two settings considered in Theorem 2. Fig. 1(b) shows that for $r = m/\sqrt{k}$ the success rate goes to zero as k increases, as predicted by the first part of Theorem 2.[3] In Fig. 1(c) we see that the success rate does not go to zero for $r = m/2$; however, it is always bounded by $1 - e^{-0.5} \approx 0.39$, as claimed by the second part of Theorem 2.

6. APPLICATION: COLUMN SUBSET SELECTION

In this section, we present results of an empirical evaluation of the proposed PSG scheme; specifically, the performance of PSG is compared to several baselines on the task of dimensionality reduction via column subset selection (CSS) [24] for sparse subspace clustering (SSC) [25, 26] .

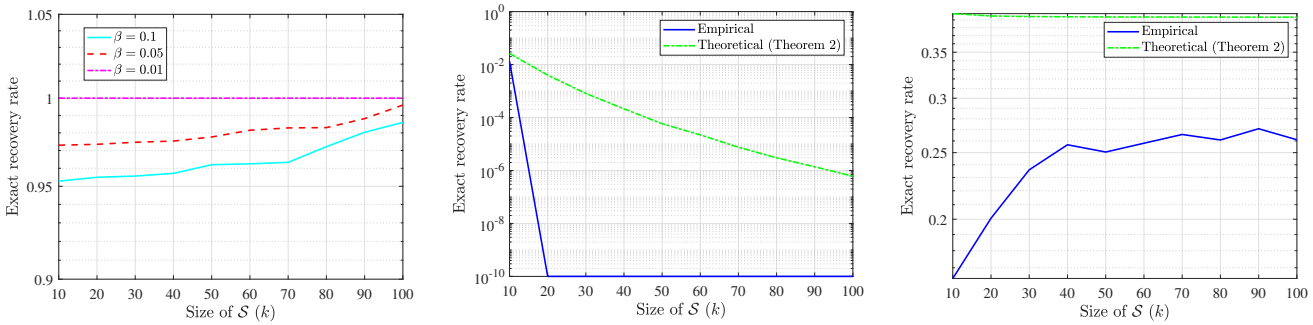
The goal of CSS is to identify a subset $\mathcal{S}$, $|\mathcal{S}| = k$, of the set of m columns of a data matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ that best approximate the entire data matrix. Formally, the task of identifying $\mathcal{S}$ can be cast as the optimization problem

$$\underset{\mathcal{S}}{\text{minimize}} \quad \|\mathbf{A} - \mathbf{P}_\mathcal{S}\mathbf{A}\|_F^2 \quad \text{s.t.} \quad |\mathcal{S}| = k, \tag{11}$$

where $\mathbf{P}_\mathcal{S} = \mathbf{A}_\mathcal{S}\mathbf{A}_\mathcal{S}^\dagger$ is the projection operator onto the span of columns of $\mathbf{A}_\mathcal{S}$ and $\mathbf{A}_\mathcal{S}^\dagger = \left(\mathbf{A}_\mathcal{S}^\top\mathbf{A}_\mathcal{S}\right)^{-1}\mathbf{A}_\mathcal{S}^\top$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{A}_\mathcal{S}$. Since $\mathbf{A} = \mathbf{P}_\mathcal{S}\mathbf{A} + (\mathbf{I} - \mathbf{P}_\mathcal{S})\mathbf{A}$ and $\|\mathbf{A}\|_F^2 = \|\mathbf{P}_\mathcal{S}\mathbf{A}\|_F^2 + \|(I - \mathbf{P}_\mathcal{S})\mathbf{A}\|_F^2$ by properties of projection

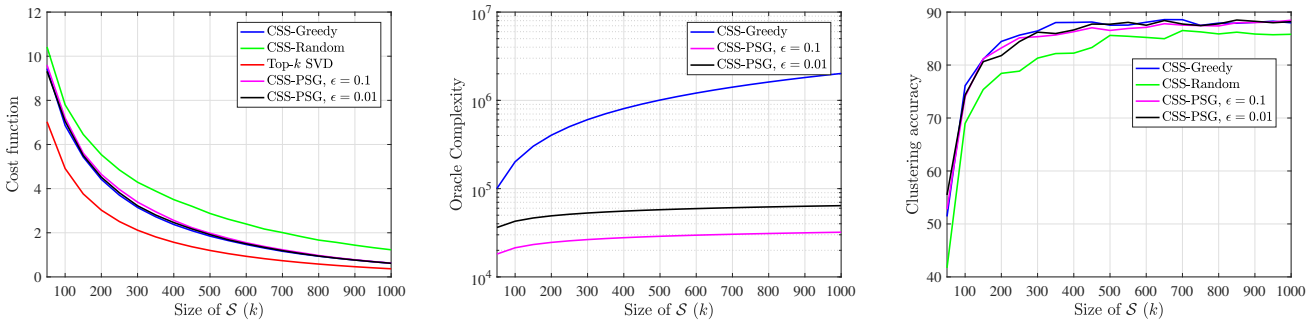
[2] Note that $\mathcal{S}^{(k)}_{alg}$ and $\mathcal{S}^\star$ are quantities that depend on m.

[3] Note that, in this setting, for $k \geq 20$ ALG failed in all of the trials; however, for illustration purposes (i.e., to be able to show the plot in the logarithmic scale) we set the success rate of ALG for $k \geq 20$ to 10^{-10}.



(a) PSG's recovery performance (b) $r = m/\sqrt{k}$ (c) $r = m/2$

Fig. 1: Empirical evaluation of the theoretical bounds established by Theorem 2.



(a) Reconstruction error (b) Oracle complexity (c) Clustering accuracy

Fig. 2: Performance comparison of SSC with various CSS-based dimensionality reduction schemes on EYaleB dataset consisting of face images under 64 different illumination conditions.

matrices, (11) can equivalently be written as an instance of the weak submodular maximization task in (2) [27, 28].

Here we aim to use Algorithm 1 as a CSS-based dimensionality reduction technique to reduce the cost of performing clustering via SSC [25, 26]. That is, using a lower dimensional data matrix $\mathbf{A}_{\mathcal{S}_g}$ obtained via CSS, we learn the representation matrix $\mathbf{C}$ by solving

$$\underset{\mathbf{C}}{\text{minimize}} \quad \|\mathbf{A}_{\mathcal{S}_g} - \mathbf{A}_{\mathcal{S}_g}\mathbf{C}\|_F^2 + \lambda\|\mathbf{C}\|_1, \tag{12}$$

and then employ spectral clustering [29] on $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^\top$ to segment the data points.

We consider the proposed PSG scheme with two values of ϵ: $\epsilon = 0.1$ and $\epsilon = 0.01$. We consider GREEDY and random column subset selection as the benchmarking schemes. Additionally, we use the best rank-k approximation of a matrix (i.e., top-k SVD) to serve as an upper bound on the achievable performance; note that this scheme explicitly minimizes the Forbenius reconstruction criteria. We compare performance of the above algorithms using the real EYaleB dataset [30] which contains frontal face images of 38 individuals under 64 different illumination conditions. There are $m = 2414$ columns (i.e., features) in this dataset; we select k out of $m = 2414$ columns, where k varies from 100 to 1000, and apply the SSC method of [26] to cluster the data points based on the selected features.

Fig. 2 shows the performance of various column subset selection schemes as well as the top-k SVD approach. In Fig. 2(a) we observe that the reconstruction errors of GREEDY and the proposed scheme are nearly identical, and that as we increase the number of selected columns the reconstruction error decreases; this is consistent with the fact that $f(\mathcal{S})$ is a monotone function. Fig. 2(b) shows a significant computational complexity improvement that the proposed scheme provides over the greedy CSS method. Since the complexity of Algorithm 1 increases logarithmically in k, the cost of selecting more columns is relatively small compared to the greedy approach. Note that we observe $\epsilon = 0.1$ achieves the best tradeoff between computational costs and performance. Furthermore, depending on the amount of data available, the value epsilon can be tuned with cross-validation. Finally, in Fig. 2(c) we compare the clustering accuracy of SSC applied to a subset of features selected by different schemes. As the figure shows, clustering performance of SSC combined with the proposed CSS method is nearly identical to that of greedy; moreover, both achieve superior accuracy compared to schemes that randomly select subsets of columns.

7. CONCLUSION

In this paper, we studied the problem of large-scale monotone weak submodular maximization that comes up in many modern signal processing and machine learning applications including sparse reconstruction, dimensionality reduction, observation gathering, and sensor selection. Motivated by the desire to reduce complexity of the celebrated greedy scheme, we theoretically studied fundamental performance limits of restricting the size of the greedy search space by means of uniform sampling strategies. We showed that an increasing schedule of the search space size satisfies a necessary condition for the exact identification of the optimal subset in large-scale problems. Following this insight, we proposed a progressive stochastic greedy algorithm and demonstrated its efficacy in the applications to sparse subset selection and dimensionality reduction.

8. REFERENCES

[1] Andreas Krause and Daniel Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*, pp. 71–104. Cambridge University Press, 2014.

[2] David P Williamson and David B Shmoys, *The design of approximation algorithms*, Cambridge university press, 2011.

[3] David Kempe, Jon Kleinberg, and Éva Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[4] Andreas Krause, Ajit Singh, and Carlos Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[5] Andrew Guillory and Jeff A Bilmes, "Active semi-supervised learning using submodular functions," *arXiv preprint arXiv:1202.3726*, 2012.

[6] Joel A Tropp and Anna C Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[7] Siddharth Joshi and Stephen Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[8] Abhimanyu Das and David Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011, pp. 1057–1064.

[9] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban, "Restricted strong convexity implies weak submodularity," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3539–3568, 2018.

[10] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher, "An analysis of approximations for maximizing submodular set functions?i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[11] Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause, "Distributed submodular maximization: Identifying representative elements in massive data," in *Advances in Neural Information Processing Systems*, 2013, pp. 2049–2057.

[12] Alina Ene and Huy L Nguyen, "Submodular maximization with nearly-optimal approximation and adaptivity in nearly-linear time," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 274–282.

[13] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause, "Lazier than lazy greedy," in *AAAI Conference on Artificial Intelligence*. AAAI, 2015.

[14] Abolfazl Hashemi, Mahsa Ghasemi, Haris Vikalo, and Ufuk Topcu, "A randomized greedy algorithm for near-optimal sensor scheduling in large-scale sensor networks," in *American Control Conference (ACC)*. IEEE, 2018, pp. 1027–1032.

[15] Uriel Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM*, vol. 45, no. 4, pp. 634–652, Jul. 1998.

[16] Haifeng Zhang and Yevgeniy Vorobeychik, "Submodular optimization with routing constraints," in *AAAI Conference on Artificial Intelligence*, 2016.

[17] Thibaut Horel and Yaron Singer, "Maximization of approximately submodular functions," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 3045–3053.

[18] Luiz Chamon and Alejandro Ribeiro, "Approximate supermodularity bounds for experimental design," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5409–5418.

[19] Rajiv Khanna, Ethan Elenberg, Alex Dimakis, Sahand Negahban, and Joydeep Ghosh, "Scalable greedy feature selection via weak submodularity," in *Artificial Intelligence and Statistics*, 2017, pp. 1560–1568.

[20] Abolfazl Hashemi, Mahsa Ghasemi, and Haris Vikalo, "Submodular observation selection and information gathering for quadratic models," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol. 97.

[21] Abolfazl Hashemi, Haris Vikalo, and Gustavo de Veciana, "Progressive stochastic greedy sparse reconstruction and support selection," *arXiv preprint arXiv:1907.09064*, 2019.

[22] Joel A Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[23] Emmanuel J Candes and Terence Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[24] Joel A Tropp, "Column subset selection, matrix factorization, and eigenvalue optimization," in *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2009, pp. 978–986.

[25] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 2790–2797.

[26] Chong You, Daniel Robinson, and René Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3918–3927.

[27] Ahmed K Farahat, Ahmed Elgohary, Ali Ghodsi, and Mohamed S Kamel, "Greedy column subset selection for large-scale data sets," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 1–34, 2015.

[28] Aditya Bhaskara, Afshin Rostamizadeh, Jason Altschuler, Morteza Zadimoghaddam, Thomas Fu, and Vahab Mirrokni, "Greedy column subset selection: New bounds and distributed algorithms," in *International Conference on Machine Learning (ICML)*. Omnipress, 2016.

[29] Andrew Y Ng, Michael I Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[30] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.