Approximation Guarantees of Stochastic Greedy Algorithms for Subset Selection

Chao Qian[1], **Yang Yu**[2], **Ke Tang**[3]
[1] Anhui Province Key Lab of Big Data Analysis and Application,
University of Science and Technology of China, Hefei 230027, China
[2] National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210023, China
[3] Shenzhen Key Lab of Computational Intelligence, Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen 518055, China
chaoqian@ustc.edu.cn, yuy@nju.edu.cn, tangk3@sustc.edu.cn

Abstract

Subset selection is a fundamental problem in many areas, which aims to select the best subset of size at most k from a universe. Greedy algorithms are widely used for subset selection, and have shown good approximation performances in deterministic situations. However, their behaviors are stochastic in many realistic situations (e.g., large-scale and noisy). For general stochastic greedy algorithms, bounded approximation guarantees were obtained only for subset selection with monotone submodular objective functions, while real-world applications often involve non-monotone or non-submodular objective functions and can be subject to a more general constraint than a size constraint. This work proves their approximation guarantees in these cases, and thus largely extends the applicability of stochastic greedy algorithms.

1 Introduction

The subset selection problem is to select a subset of size at most k from a ground set of n items for maximizing some given objective function f. It arises in many applications, such as sparse regression [Miller, 2002], influence maximization [Kempe *et al.*, 2003], and sensor placement [Krause *et al.*, 2008], to name a few. For this general NP-hard problem, greedy algorithms were shown to be powerful. For example, when the objective function f satisfies the monotone and submodular property, the standard greedy algorithm, which iteratively adds one item with the largest marginal gain on f, achieves the optimal approximation guarantee of $(1 - 1/e)$ [Nemhauser *et al.*, 1978; Nemhauser and Wolsey, 1978].

However, greedy algorithms can only be performed stochastically in many realistic situations. For example, when the ground set is very large or the objective function evaluation is noisy, the standard greedy algorithm can only select an item whose marginal gain is approximately optimal in expectation at each greedy step. For large-scale applications, the random sampling technique is often employed to improve the scalability [Mirzasoleiman *et al.*, 2015; Ohsaka and Yoshida, 2015], which makes that the best item from a random subset (instead of the whole set) of remaining items is selected in each iteration. For noisy applications, the noise in the objective function evaluation makes that the item with the largest noisy marginal gain (instead of the largest true marginal gain) is selected in each iteration [Singla *et al.*, 2016; Qian *et al.*, 2017b].

To the best of our knowledge, general stochastic greedy algorithms for subset selection were not studied until recently. Hassidim and Singer [2017] considered a general framework for stochastic variants of the standard greedy algorithm, called STOCHASTIC-STANDARD-GREEDY. In each iteration, a value ξ is randomly sampled from a distribution $\mathcal{D}$, and then the algorithm selects one item whose marginal gain is a ξ-approximation of the largest marginal gain (i.e., at least the largest marginal gain times a factor of ξ). They proved that when f is monotone submodular, STOCHASTIC-STANDARD-GREEDY achieves a $(1 - e^{-\mu})$-approximation guarantee, where μ is the expectation of $\xi \sim \mathcal{D}$.

Note that real-world applications of subset selection often involve non-submodular or non-monotone objective functions, and can be subject to a more general constraint than a size constraint. In this paper, we thus theoretically study the approximation performance of the stochastic version of the corresponding greedy algorithms in these cases. Our analysis uses some quantities γ, λ and α, which characterize how close an arbitrary function f is to submodularity from different aspects [Das and Kempe, 2011; Zhou and Spanos, 2016; Zhang and Vorobeychik, 2016]. The main contributions are:

- For subset selection with a monotone (not necessarily submodular) objective function and a size constraint, the STOCHASTIC-STANDARD-GREEDY algorithm obtains an approximation guarantee of $(1 - e^{-\mu\gamma})$ (**Theorem 1**).
- For subset selection with a non-monotone (not necessarily submodular) objective function and a size constraint, the STOCHASTIC-RANDOM-GREEDY algorithm obtains an approximation guarantee of $\frac{\mu}{e}$ plus a term depending on λ (**Theorem 2**).

- For subset selection with a monotone (not necessarily submodular) objective function and a p-system constraint, the STOCHASTIC-GENERAL-GREEDY algorithm obtains an approximation guarantee of $\frac{\alpha^2\mu}{p+\alpha^2\mu}$ (**Theorem 3**).

We also show how to bound the quantities γ, λ and α in several real-world applications, implying the practical applicability of our derived approximation guarantees.

2 Preliminaries

Let $\mathbb{R}$ and $\mathbb{R}^+$ denote the set of reals and non-negative reals, respectively. Given a ground set $V = \{v_1, v_2, \ldots, v_n\}$, we study the functions $f : 2^V \to \mathbb{R}$ over subsets of V. A set function f is monotone if for any $S \subseteq T$, $f(S) \leq f(T)$. Without loss of generality, we assume that monotone functions are normalized, i.e., $f(\emptyset) = 0$. Let $f_S(T)$ denote the marginal value $f(S \cup T) - f(S)$. A set function $f : 2^V \to \mathbb{R}$ is submodular [Nemhauser *et al.*, 1978] if for any $S \subseteq T \subseteq V$,

$$f(T) - f(S) \leq \sum\nolimits_{v \in T \setminus S} f_S(v), \tag{1}$$

or equivalently, for any $S \subseteq T \subseteq V$ and $v \notin T$,

$$f_S(v) \geq f_T(v). \tag{2}$$

Note that we represent a singleton set $\{v\}$ by v for simplicity. We then give three notions of "approximate submodularity", which measure to what extent a general set function f has the submodular property. The γ-submodularity ratio and the λ-submodularity index are defined based on Eq. (1), while the α-submodularity ratio is defined based on Eq. (2). It is easy to verify that $\gamma_{S,k}(f) \leq 1$, $\lambda_{S,k}(f) \leq 0$ and $\alpha_f \leq 1$. When f is clear, we will use $\gamma_{S,k}$, $\lambda_{S,k}$ and α for short.

Definition 1 (γ-Submodularity Ratio [Das and Kempe, 2011])**.** *The submodularity ratio of a set function $f : 2^V \to \mathbb{R}$ with respect to a set S and a parameter $k \geq 1$ is*

$$\gamma_{S,k}(f) = \min_{L \subseteq S, T : |T| \leq k, T \cap L = \emptyset} \frac{\sum_{v \in T} f_L(v)}{f_L(T)}.$$

Definition 2 (λ-Submodularity Index [Zhou and Spanos, 2016])**.** *The submodularity index of a set function $f : 2^V \to \mathbb{R}$ with respect to a set S and a parameter $k \geq 1$ is*

$$\lambda_{S,k}(f) = \min_{L \subseteq S, T : |T| \leq k, T \cap L = \emptyset} \sum\nolimits_{v \in T} f_L(v) - f_L(T).$$

Definition 3 (α-Submodularity Ratio [Zhang and Vorobeychik, 2016; Qian *et al.*, 2017a])**.** *The submodularity ratio of a set function $f : 2^V \to \mathbb{R}$ is*

$$\alpha_f = \min_{S \subseteq T, v \notin T} \frac{f_S(v)}{f_T(v)}.$$

Remark 1. *For a general set function f, f is submodular iff $\lambda_{S,k}(f) = 0$ for any S and k. For a monotone set function f, it holds: (1) f is submodular iff $\gamma_{S,k}(f) = 1$ for any S and k; (2) f is submodular iff $\alpha_f = 1$; (3) $\alpha_f \leq \gamma_{S,k}(f)$ for any S and k, since $f_L(T)$ in Definition 1 can be upper bounded by*

$$f_L(T) = \sum_{i=1}^{|T|} f_{L \cup \{v_1^*, \ldots, v_{i-1}^*\}}(v_i^*) \leq \sum_{i=1}^{|T|} \frac{f_L(v_i^*)}{\alpha_f} = \frac{\sum_{v \in T} f_L(v)}{\alpha_f},$$

where the items in T are denoted as $\{v_1^*, \ldots, v_{|T|}^*\}$ and $|\cdot|$ denotes the size of a set.

Algorithm 1 STOCHASTIC-STANDARD-GREEDY Algorithm

Input: a budget k and a distribution $\mathcal{D}$
Output: a subset of V with k items
Process:

1: Let $S = \emptyset$.
2: **repeat**
3: $\quad \xi \leftarrow$ randomly sampled from $\mathcal{D}$.
4: $\quad v^* \leftarrow$ an arbitrary item from $V \setminus S$ s.t. $f_S(v^*) \geq \xi \cdot \max_{v \in V \setminus S} f_S(v)$.
5: $\quad S \leftarrow S \cup v^*$.
6: **until** $|S| = k$
7: **return** S

An independence system is a pair $(V, \mathcal{I})$, where V is a ground set and $\mathcal{I} \subseteq 2^V$, satisfying that $\emptyset \in \mathcal{I}$ and $\forall S \subseteq T \in \mathcal{I} : S \in \mathcal{I}$. Any set $S \in \mathcal{I}$ is called an independent set. For any $S \subseteq V$, a maximal independent subset of S is called a basis of S. The general subset selection problem is presented in Definition 4, which is to find an independent subset of V maximizing a given objective function f. Without loss of generality, we assume that f is non-negative.

Definition 4 (Subset Selection)**.** *Given all items $V = \{v_1, v_2, \ldots, v_n\}$, an objective function $f : 2^V \to \mathbb{R}^+$ and an independence system $(V, \mathcal{I})$, the goal is to find a subset $S \subseteq V$ maximizing $f(S)$ such that $S \in \mathcal{I}$, i.e.,*

$$\arg\max_{S \subseteq V} f(S) \quad \textit{s.t.} \quad S \in \mathcal{I}.$$

In this paper, we will study the subset selection problem with two specific independence systems, a uniform matroid and a p-system. An independence system $(V, \mathcal{I})$ is a matroid if it satisfies the additional property: $\forall S, T \in \mathcal{I}, |S| > |T| : \exists v \in S \setminus T, T \cup v \in \mathcal{I}$. A uniform matroid is $(V, \mathcal{I})$ with $\mathcal{I} = \{S \subseteq V \mid |S| \leq k\}$, which is actually a size constraint $|S| \leq k$. An independence system $(V, \mathcal{I})$ is a p-system [Jenkyns, 1976; Korte and Hausmann, 1978], if for any $S \subseteq V$,

$$\frac{\max_{L : L \text{ is a basis of } S} |L|}{\min_{L : L \text{ is a basis of } S} |L|} \leq p. \tag{3}$$

Note that a p-system is much more general than a uniform matroid, and even covers the intersection of p matroids (i.e., $(V, \cap_{i=1}^p \mathcal{I}_i)$, where each $(V, \mathcal{I}_i)$ is a matroid) [Calinescu *et al.*, 2011].

3 Non-submodular Functions

For the subset selection problem with a monotone submodular objective function f and a size constraint $|S| \leq k$, it was proved that STOCHASTIC-STANDARD-GREEDY finds a subset S of V with $\mathbb{E}[f(S)] \geq (1 - e^{-\mu}) \cdot OPT$ [Hassidim and Singer, 2017]. Note that $\mathbb{E}[\cdot]$ denotes the expectation of a random variable, and OPT denotes the optimal function value. As presented in Algorithm 1, instead of selecting one item with the largest marginal gain, the stochastic algorithm selects one item whose marginal gain is at least a factor of ξ from the largest marginal gain in each iteration, where $\xi \in (0, 1]$ is drawn i.i.d. from some distribution $\mathcal{D}$. Let μ denote the expectation of $\xi \sim \mathcal{D}$.

In this section, we consider a more general situation, where f is not necessarily submodular. The approximation guarantee is shown in Theorem 1. The proof relies on Lemma 1, which gives the expected improvement on f in one iteration.

Lemma 1. *Let S_i denote the subset S after i iterations of* STOCHASTIC-STANDARD-GREEDY. *Then, we have*

$$\mathbb{E}[f(S_{i+1}) - f(S_i) \mid S_i] \geq \frac{\mu\gamma_{S_i,k}}{k} \cdot (OPT - f(S_i)).$$

Proof. Let S^* be an optimal subset, i.e., $f(S^*) = OPT$, and let ξ_i denote the value of ξ in the i-th iteration of Algorithm 1. Then, for any given S_i, we have

$$f(S_{i+1}) - f(S_i) \geq \xi_{i+1} \max_{v \in V \setminus S_i} f_{S_i}(v) \geq \frac{\xi_{i+1}}{k} \sum_{v \in S^* \setminus S_i} f_{S_i}(v)$$

$$\geq \frac{\xi_{i+1}\gamma_{S_i,k}}{k} f_{S_i}(S^* \setminus S_i) \geq \frac{\xi_{i+1}\gamma_{S_i,k}}{k}(OPT - f(S_i)),$$

where the first inequality is by lines 3-5 of Algorithm 1, the second is by $|S^* \setminus S_i| \leq k$, the third is by Definition 1 and the monotonicity of f, and the last is by $f(S^* \cup S_i) \geq f(S^*) = OPT$. Taking the expectation over both sides, the lemma holds, since $\mathbb{E}[\xi_{i+1} \mid \xi_{i+1} \sim \mathcal{D}] = \mu$. □

Theorem 1. *For subset selection with a monotone objective function and a size constraint,* STOCHASTIC-STANDARD-GREEDY *finds a subset $S \subseteq V$ with $|S| = k$ and*

$$\mathbb{E}[f(S)] \geq (1 - e^{-\mu\gamma_{\min}}) \cdot OPT,$$

where $\gamma_{\min} = \min_{S:|S|=k-1} \gamma_{S,k}$.

Proof. Taking the expectation over S_i, we know from Lemma 1 that for $0 \leq i \leq k-1$,

$$\mathbb{E}[f(S_{i+1})] - \mathbb{E}[f(S_i)] \geq \frac{\mu\gamma_{\min}}{k} \cdot (OPT - \mathbb{E}[f(S_i)]).$$

Note that $\gamma_{S_i,k} \geq \gamma_{\min}$ since $|S_i| \leq k-1$ and $\gamma_{S,k}$ decreases with S. By a simple transformation, we can equivalently get

$$\mathbb{E}[f(S_{i+1})] \geq \left(1 - \frac{\mu\gamma_{\min}}{k}\right) \mathbb{E}[f(S_i)] + \frac{\mu\gamma_{\min}}{k} OPT.$$

Based on this inequality, an inductive proof can show that, for $0 \leq i \leq k$,

$$\mathbb{E}[f(S_i)] \geq \left(1 - \left(1 - \frac{\mu\gamma_{\min}}{k}\right)^i\right) \cdot OPT.$$

Thus, for the returned subset S_k, we get

$$\mathbb{E}[f(S_k)] \geq \left(1 - \left(1 - \frac{\mu\gamma_{\min}}{k}\right)^k\right) \cdot OPT$$
$$\geq (1 - e^{-\mu\gamma_{\min}}) \cdot OPT.$$ □

Note that our derived approximation guarantee in Theorem 1 is consistent with known results in specific cases.

Remark 2. *When f is submodular (where $\gamma_{\min} = 1$), it recovers the approximation ratio $1 - e^{-\mu}$ [Hassidim and Singer, 2017]; when the stochastic behavior is due to random sampling (where $\mu \geq 1 - \delta$), it recovers the approximation ratio $1 - e^{-(1-\delta)\gamma_{\min}}$ [Khanna* et al.*, 2017]; when the algorithm performs exactly (where $\mu = 1$), it recovers the approximation ratio $1 - e^{-\gamma_{S,k}}$ [Das and Kempe, 2011], where S is the returned subset.*

Algorithm 2 STOCHASTIC-RANDOM-GREEDY Algorithm

Input: a budget k and a distribution $\mathcal{D}$
Output: a subset of V with k items
Process:

1: Let $S = \emptyset$.
2: **repeat**
3: $\quad \xi \leftarrow$ randomly sampled from $\mathcal{D}$.
4: $\quad U^* \leftarrow$ an arbitrary subset of $V \setminus S$ with size k s.t. $\sum_{v \in U^*} f_S(v) \geq \xi \cdot \max_{U \subseteq V \setminus S, |U|=k} \sum_{v \in U} f_S(v)$.
5: $\quad v \leftarrow$ uniformly chosen from U^* at random.
6: $\quad S \leftarrow S \cup v$.
7: **until** $|S| = k$
8: **return** S

4 Non-monotone Functions

In this section, we consider the subset selection problem with a non-monotone (not necessarily submodular) objective function f and a size constraint. It was known that when f is submodular, the standard greedy algorithm fails to provide any constant guarantee, while the random greedy algorithm can achieve a $(1/e)$-approximation guarantee [Buchbinder *et al.*, 2014]. Instead of selecting the best item in each iteration, the random greedy algorithm selects one from the best k items uniformly at random. We thus analyze the stochastic version of the random greedy algorithm, called STOCHASTIC-RANDOM-GREEDY as presented in Algorithm 2. As in [Buchbinder *et al.*, 2014], we also assume that there are $2k$ "dummy" items in V whose marginal gain to any set is 0. We prove the approximation guarantee in Theorem 2 by utilizing the recursive inequality shown in Lemma 3. Note that the submodularity index λ is used here instead of the submodularity ratio γ, since γ can be negative for a non-monotone function f.

Lemma 2. *[Zhou and Spanos, 2016] Given a set function $f : 2^V \to \mathbb{R}^+$, let $S(p)$ be a random subset of $S \subseteq V$, where each item of S appears in $S(p)$ with probability at most p. Then, we have*

$$\mathbb{E}[f(S(p))] \geq (1-p)f(\emptyset) + \frac{|S|(|S|-1)}{2}\lambda_{S,2}.$$

Lemma 3. *Let S_i denote the subset S after i iterations of* STOCHASTIC-RANDOM-GREEDY, *and let S^* denote an optimal subset. Then, we have*

$$\mathbb{E}[f(S_{i+1}) - f(S_i) \mid S_i] \geq \frac{\mu}{k}(f(S_i \cup S^*) - f(S_i) + \lambda_{S_i,k}).$$

Proof. According to lines 3-6 of Algorithm 2 and $\mathbb{E}[\xi \mid \xi \sim \mathcal{D}] = \mu$, we have

$$\mathbb{E}[f(S_{i+1}) - f(S_i) \mid S_i] \geq \frac{\mu}{k} \max_{U \subseteq V \setminus S_i, |U|=k} \sum_{v \in U} f_{S_i}(v)$$
$$\geq \frac{\mu}{k} \sum_{v \in S^* \setminus S_i} f_{S_i}(v) \geq \frac{\mu}{k}(f_{S_i}(S^* \setminus S_i) + \lambda_{S_i,k})$$
$$= \frac{\mu}{k}(f(S_i \cup S^*) - f(S_i) + \lambda_{S_i,k}),$$

where the second inequality is by $|S^* \setminus S_i| \leq k$ and the existence of "dummy" items, and the last is by Definition 2. □

Theorem 2. *For subset selection with a non-monotone objective function and a size constraint,* STOCHASTIC-RANDOM-GREEDY *finds a subset $S \subseteq V$ with $|S| = k$ and*

$$\mathbb{E}[f(S)] \geq \frac{\mu}{e} \cdot OPT + \mu \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right),$$

where $\lambda_{\min} = \min_{S:|S|=k-1} \lambda_{S,k}$.

Proof. Taking the expectation over S_i, we know from Lemma 3 that for $0 \leq i \leq k-1$,

$$\begin{aligned} &\mathbb{E}[f(S_{i+1})] - \mathbb{E}[f(S_i)] \\ &\geq \frac{\mu}{k} \cdot (\mathbb{E}[f(S_i \cup S^*)] - \mathbb{E}[f(S_i)] + \lambda_{\min}). \end{aligned}$$

Note that $\lambda_{S_i,k} \geq \lambda_{\min}$ since $|S_i| \leq k-1$ and $\lambda_{S,k}$ decreases with S. We define a function $g : 2^{V \setminus S^*} \to \mathbb{R}^+$ as for any $S \subseteq V \setminus S^*$, $g(S) = f(S \cup S^*)$. Then,

$$\mathbb{E}[f(S_i \cup S^*)] = \mathbb{E}[g(S_i \setminus S^*)].$$

It is easy to see that any item is selected with probability at most $1/k$ in each iteration of Algorithm 2. Thus, any item of $V \setminus S^*$ appears in $S_i \setminus S^*$ with probability at most $1 - (1 - 1/k)^i$. Note that $S_i \setminus S^*$ is a random subset of $V \setminus S^*$. By applying Lemma 2 to the function g, we get

$$\begin{aligned} &\mathbb{E}[g(S_i \setminus S^*)] \\ &\geq \left(1 - \frac{1}{k}\right)^i g(\emptyset) + \frac{|V \setminus S^*|(|V \setminus S^*| - 1)}{2} \lambda_{V \setminus S^*,2}(g) \\ &\geq \left(1 - \frac{1}{k}\right)^i OPT + \frac{n(n-1)}{2} \lambda_{V \setminus S^*,2}(g), \end{aligned}$$

where the last inequality is by $g(\emptyset) = f(S^*) = OPT$, $|V \setminus S^*| \leq n$ and $\lambda_{V \setminus S^*,2}(g) \leq 0$. Based on Definition 2, we get

$$\begin{aligned} &\lambda_{V \setminus S^*,2}(g) = \min_{L \subseteq V \setminus S^*, T:|T| \leq 2, T \cap L = \emptyset} \sum_{v \in T} g_L(v) - g_L(T) \\ &= \min_{L \subseteq V \setminus S^*, T:|T| \leq 2, T \cap L = \emptyset} \sum_{v \in T} f_{L \cup S^*}(v) - f_{L \cup S^*}(T) \\ &\geq \min_{L \subseteq V, T:|T| \leq 2, T \cap L = \emptyset} \sum_{v \in T} f_L(v) - f_L(T) = \lambda_{V,2}(f). \end{aligned}$$

Thus, we get

$$\begin{aligned} &\mathbb{E}[f(S_{i+1})] - \left(1 - \frac{\mu}{k}\right) \mathbb{E}[f(S_i)] \\ &\geq \frac{\mu}{k} \cdot \left(\left(1 - \frac{1}{k}\right)^i OPT + \frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right). \end{aligned}$$

Based on this recursive inequality, an inductive proof can show that for $0 \leq i \leq k$, if $\mu < 1$,

$$\begin{aligned} \mathbb{E}[f(S_i)] \geq &\left(\left(1 - \frac{\mu}{k}\right)^i - \left(1 - \frac{1}{k}\right)^i \right) \frac{\mu}{1 - \mu} \cdot OPT \\ &+ \left(1 - \left(1 - \frac{\mu}{k}\right)^i\right) \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right); \end{aligned}$$

if $\mu = 1$,

$$\begin{aligned} \mathbb{E}[f(S_i)] \geq &\frac{i}{k} \left(1 - \frac{1}{k}\right)^{i-1} \cdot OPT \\ &+ \left(1 - \left(1 - \frac{1}{k}\right)^i\right) \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right). \end{aligned}$$

Thus, for the returned subset S_k, we get, if $\mu < 1$,

$$\begin{aligned} \mathbb{E}[f(S_k)] &\geq \left(\left(1 - \frac{\mu}{k}\right)^k - \left(1 - \frac{1}{k}\right)^k \right) \frac{\mu}{1 - \mu} \cdot OPT \\ &\quad + \left(1 - \left(1 - \frac{\mu}{k}\right)^k\right) \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right) \\ &= \left(1 - \frac{1}{k}\right)^k \left(\left(1 + \frac{1-\mu}{k-1}\right)^k - 1 \right) \frac{\mu}{1-\mu} \cdot OPT \\ &\quad + \left(1 - \left(1 - \frac{\mu}{k}\right)^k\right) \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right) \\ &\geq \frac{\mu}{e} \cdot OPT + \mu \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right), \end{aligned}$$

where the last inequality is derived by $(1 + \frac{1-\mu}{k-1})^k \geq 1 + \frac{(1-\mu)k}{k-1}$, $(1 - \frac{1}{k})^{k-1} \geq \frac{1}{e}$, $(1 - \frac{\mu}{k})^k \geq 1 - \mu$ and $\lambda_{S,k} \leq 0$ for any $S \subseteq V$ and $k \geq 1$. If $\mu = 1$, we have

$$\begin{aligned} \mathbb{E}[f(S_k)] &\geq \frac{k}{k} \left(1 - \frac{1}{k}\right)^{k-1} \cdot OPT \\ &\quad + \left(1 - \left(1 - \frac{1}{k}\right)^k\right) \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right) \\ &\geq \frac{1}{e} \cdot OPT + \frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min}. \end{aligned}$$

Thus, we can get a unified lower bound on $\mathbb{E}[f(S_k)]$ for any $\mu \in (0, 1]$, i.e.,

$$\mathbb{E}[f(S_k)] \geq \frac{\mu}{e} \cdot OPT + \mu \left(\frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min} \right). \quad \square$$

Note that this derived approximation guarantee is consistent with known results for the random greedy algorithm.

Remark 3. *When the algorithm performs exactly (where $\mu = 1$), it recovers the approximation bound $\frac{1}{e} \cdot OPT + \frac{n(n-1)}{2} \lambda_{V,2} + \lambda_{\min}$ [Zhou and Spanos, 2016], which further recovers the approximation ratio $\frac{1}{e}$ for submodular f (where $\lambda_{S,k} = 0$ for any $S \subseteq V$ and $k \geq 1$) [Buchbinder* et al.*, 2014].*

5 General Constraints

In this section, we consider a more general constraint, i.e., a subset S belongs to a p-system. Note that a p-system covers an intersection of p matroids, and of course covers a uniform matroid (i.e., a size constraint $|S| \leq k$). That is, we study the subset selection problem with a monotone (not necessarily submodular) objective function f and a p-system constraint. It was known that when f is submodular, the general greedy algorithm, which iteratively adds one

Algorithm 3 STOCHASTIC-GENERAL-GREEDY Algorithm

Input: an independence system $(V, \mathcal{I})$ and a distribution $\mathcal{D}$
Output: a basis of V
Process:

1: Let $S = \emptyset$ and $U = V$.
2: **repeat**
3: $\quad U \leftarrow \{v \in U \mid S \cup v \in \mathcal{I}\}$
4: $\quad$ **if** $U \neq \emptyset$ **then**
5: $\quad\quad \xi \leftarrow$ randomly sampled from $\mathcal{D}$.
6: $\quad\quad v^* \leftarrow$ an arbitrary item from U s.t. $f_S(v^*) \geq \xi \cdot \max_{v \in U} f_S(v)$.
7: $\quad\quad S \leftarrow S \cup v^*$ and $U \leftarrow U \setminus v^*$.
8: $\quad$ **end if**
9: **until** $U = \emptyset$
10: **return** S

item with the largest marginal gain among those items that keep the set independent, can achieve a tight approximation guarantee of $1/(p+1)$ [Calinescu *et al.*, 2011]. We thus analyze the stochastic version of the general greedy algorithm, called STOCHASTIC-GENERAL-GREEDY as presented in Algorithm 3. We prove the approximation guarantee in Theorem 3. Note that the submodularity ratio α is used here, since we need a weak version of the diminishing return property (i.e., Eq. (2)) in the proof.

Lemma 4. *[Fisher* et al.*, 1978] For $\delta_i, \rho_i \geq 0$ with $0 \leq i \leq k-1$, if it satisfies that $\sum_{i=0}^{t-1} \delta_i \leq t$ for $1 \leq t \leq k$ and $\rho_{i-1} \geq \rho_i$ for $1 \leq i \leq k-1$, then $\sum_{i=0}^{k-1} \delta_i \rho_i \leq \sum_{i=0}^{k-1} \rho_i$.*

Theorem 3. *For subset selection with a monotone objective function and a p-system constraint,* STOCHASTIC-GENERAL-GREEDY *finds a basis S of V with*

$$\mathbb{E}[f(S)] \geq \frac{\alpha^2 \mu}{p + \alpha^2 \mu} \cdot OPT.$$

Proof. Let S^* denote an optimal subset, i.e., $f(S^*) = OPT$. Assume that the returned basis by STOCHASTIC-GENERAL-GREEDY (i.e., Algorithm 3) contains k items. Let S_i $(0 \leq i \leq k)$ denote the subset S after i iterations of Algorithm 3, where $S_0 = \emptyset$ and S_k is the returned basis. By Definition 1 and the monotonicity of f, we have

$$\sum_{v \in S^* \setminus S_k} f_{S_k}(v) \geq \gamma_{S_k, |S^* \setminus S_k|} \cdot (f(S_k \cup S^*) - f(S_k)) \quad (4)$$
$$\geq \gamma_{S_k, |S^* \setminus S_k|} \cdot (OPT - f(S_k)).$$

For $0 \leq i \leq k$, let $X_i = S_i \cup \{v \in V \mid S_i \cup v \notin \mathcal{I}\}$. Then, in the $(i+1)$-th iteration of Algorithm 3, the set U in line 3 is actually $V \setminus X_i$, which is the set of items whose inclusion into S_i keep the set independent. It is easy to verify that $X_i \subseteq X_{i+1}$ and $X_k = V$ since S_k is a basis (i.e., a maximal independent subset of V). By the definition of a p-system (i.e., Eq. (3)), we get, for $0 \leq i \leq k$,

$$|X_i \cap S^*| \leq p \cdot |S_i| = ip, \quad (5)$$

since S_i is a basis of X_i and $X_i \cap S^*$ is an independent subset of X_i. For $0 \leq i \leq k-1$, let $S_i^* = (X_{i+1} \setminus X_i) \cap S^*$. Since $X_i \subseteq X_{i+1}$ and $X_k = V$, we have $S_i^* \cap S_j^* = \emptyset$ for any $i \neq j$ and $\cup_{i=0}^{k-1} S_i^* = (X_k \setminus X_0) \cap S^* = (V \setminus X_0) \cap S^* = S^*$. Note that $X_0 \cap S^* = \emptyset$. That is, $\{S_0^*, S_1^*, \ldots, S_{k-1}^*\}$ is a partition of S^*. Then, we have

$$\sum_{v \in S^* \setminus S_k} f_{S_k}(v) = \sum_{v \in S^*} f_{S_k}(v) = \sum_{i=0}^{k-1} \sum_{v \in S_i^*} f_{S_k}(v).$$

Let $u_i^* \in \arg\max_{v \in V \setminus X_i} f_{S_k}(v)$. Since $S_i^* \subseteq V \setminus X_i$, we have, for any $v \in S_i^*$, $f_{S_k}(v) \leq f_{S_k}(u_i^*)$. Thus, we get

$$\sum_{v \in S^* \setminus S_k} f_{S_k}(v) \leq \sum_{i=0}^{k-1} \sum_{v \in S_i^*} f_{S_k}(u_i^*) = \sum_{i=0}^{k-1} |S_i^*| f_{S_k}(u_i^*). \quad (6)$$

Since $X_i \subseteq X_{i+1}$, we have $f_{S_k}(u_i^*) \geq f_{S_k}(u_{i+1}^*)$. For any $1 \leq t \leq k$, it holds that

$$\sum\nolimits_{i=0}^{t-1} |S_i^*|/p = |X_t \cap S^*|/p \leq t,$$

where the inequality is by Eq. (5). Thus, by Lemma 4, we get

$$\sum\nolimits_{i=0}^{k-1} (|S_i^*|/p) \cdot f_{S_k}(u_i^*) \leq \sum\nolimits_{i=0}^{k-1} f_{S_k}(u_i^*). \quad (7)$$

Since $S_i \subseteq S_k$ for any $i < k$, by Definition 3 and the monotonicity of f, we can get

$$\sum\nolimits_{i=0}^{k-1} f_{S_k}(u_i^*) \leq \sum\nolimits_{i=0}^{k-1} f_{S_i}(u_i^*)/\alpha. \quad (8)$$

Let $v_i^* \in \arg\max_{v \in V \setminus X_i} f_{S_i}(v)$. Since $u_i^* \in V \setminus X_i$, it holds

$$f_{S_i}(u_i^*) \leq f_{S_i}(v_i^*). \quad (9)$$

By applying Eqs. (7), (8) and (9) to Eq. (6), we get

$$\sum_{v \in S^* \setminus S_k} f_{S_k}(v) \leq \frac{p}{\alpha} \sum_{i=0}^{k-1} f_{S_i}(v_i^*). \quad (10)$$

By the procedure of Algorithm 3, we have

$$\mathbb{E}[f(S_{i+1}) - f(S_i) \mid S_i] \geq \mu \max_{v \in V \setminus X_i} f_{S_i}(v) = \mu f_{S_i}(v_i^*).$$

Note that $V \setminus X_i$ is the set of items whose inclusion into S_i can keep the set independent, and v_i^* is the item with the largest marginal gain among $V \setminus X_i$. Taking the expectation over S_i,

$$\mathbb{E}[f(S_{i+1}) - f(S_i)] \geq \mu \cdot \mathbb{E}[f_{S_i}(v_i^*)].$$

Then, we have

$$\mathbb{E}[f(S_k)] = \sum_{i=0}^{k-1} \mathbb{E}[f(S_{i+1}) - f(S_i)] \geq \mu \cdot \sum_{i=0}^{k-1} \mathbb{E}[f_{S_i}(v_i^*)]. \quad (11)$$

By combining Eqs. (4) and (10), taking the expectation over both sides, and using $\gamma_{S_k, |S^* \setminus S_k|} \geq \alpha$ (see Remark 1), we get

$$\alpha(OPT - \mathbb{E}[f(S_k)]) \leq \frac{p}{\alpha} \sum_{i=0}^{k-1} \mathbb{E}[f_{S_i}(v_i^*)] \leq \frac{p}{\alpha\mu} \mathbb{E}[f(S_k)],$$

where the last inequality is by Eq. (11). Thus, for the returned subset S_k, we have

$$\mathbb{E}[f(S_k)] \geq \frac{\alpha^2 \mu}{p + \alpha^2 \mu} \cdot OPT.$$

□

Note that our derived approximation guarantee in Theorem 3 is consistent with known results for the problem with submodular f and matroid constraints.

Remark 4. *When f is submodular (where $\alpha = 1$) and a p-system is specialized as an intersection of p-matroids, it recovers the approximation ratio $\frac{\mu}{p+\mu}$ [Hassidim and Singer, 2017], which further recovers the approximation ratio $\frac{1}{p+1}$ of the exact algorithm (where $\mu = 1$) [Fisher* et al.*, 1978].*

6 Applications of Approximation Guarantees

To understand the derived approximation guarantees of stochastic greedy algorithms in real-world applications, we need to provide lower bounds on γ, λ or α for the corresponding objective functions. Note that, lower bounds on γ were derived for some monotone non-submodular applications [Das and Kempe, 2011; Elenberg *et al.*, 2016; Bian *et al.*, 2017], and that on λ were also derived for the non-monotone non-submodular application of causal covariate selection [Zhou and Spanos, 2016]. In this paper, we thus only analyze the submodularity ratio α, which was never touched before. We give lower bounds on α in Lemmas 5 and 6 for the monotone non-submodular objective functions in the applications of Bayesian experimental design and non-parametric learning. The proofs are inspired from that of Propositions 1 and 2 in [Bian *et al.*, 2017], which prove lower bounds on γ.

In Bayesian experimental design, the goal is to select observations to maximize the quality of parameter estimation. Krause *et al.* [2008] considered the Bayesian A-optimality objective function, which is to maximally reduce the variance of the posterior distribution over parameters in linear models. Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ denote the observation matrix, where $\boldsymbol{x}_i \in \mathbb{R}^d$. Assume that each $\boldsymbol{x}_i$ is normalized, i.e., $\|\boldsymbol{x}_i\| = 1$. Let $X_S \in \mathbb{R}^{d\times|S|}$ denote the submatrix of X with its columns indexed by $S \subseteq \{1, 2, \ldots, n\}$. Let $\mathrm{tr}(\cdot)$ denote the trace of a matrix and let $\sigma_i(\cdot)$ denote the i-th largest singular value of a matrix. The linear model is described as $\boldsymbol{y}_S = X_S^T\boldsymbol{\theta} + \boldsymbol{w}$, where $\boldsymbol{\theta} \sim \mathcal{N}(0, \Lambda^{-1})$, $\Lambda = \beta^2 I_d$, $\boldsymbol{w} \sim \mathcal{N}(0, \delta^2 I_{|S|})$, and I_j denotes the identity matrix of size j. Then, the A-optimality objective function is defined as

$$f(S) = \mathrm{tr}(\Lambda^{-1}) - \mathrm{tr}((\Lambda + \delta^{-2}X_S X_S^T)^{-1}), \quad (12)$$

which is monotone non-submodular [Krause *et al.*, 2008].

Lemma 5. *For the A-optimality objective function (i.e., Eq. (*12*)) in Bayesian experimental design, the submodularity ratio α can be lower bounded as*

$$\alpha \geq \frac{1}{(1 + \delta^{-2}\sigma_1^2(X)/\beta^2)^2}.$$

Proof. For any column index set $S \subseteq \{1, \ldots, n\}$ and $v \notin S$,

$$\begin{aligned}
f_S(v) &= f(S \cup v) - f(S)\\
&= \mathrm{tr}((\Lambda+\delta^{-2}X_S X_S^T)^{-1}) - \mathrm{tr}((\Lambda+\delta^{-2}X_{S\cup v}X_{S\cup v}^T)^{-1})\\
&= \sum\nolimits_{i=1}^{d} \frac{1}{\beta^2+\delta^{-2}\sigma_i^2(X_S)} - \sum\nolimits_{i=1}^{d} \frac{1}{\beta^2+\delta^{-2}\sigma_i^2(X_{S\cup v})}\\
&= \sum\nolimits_{i=1}^{d} \frac{\delta^{-2}(\sigma_i^2(X_{S\cup v}) - \sigma_i^2(X_S))}{(\beta^2 + \delta^{-2}\sigma_i^2(X_S))(\beta^2 + \delta^{-2}\sigma_i^2(X_{S\cup v}))}\\
&\geq \sum\nolimits_{i=1}^{d} \frac{\delta^{-2}(\sigma_i^2(X_{S\cup v}) - \sigma_i^2(X_S))}{(\beta^2 + \delta^{-2}\sigma_1^2(X))^2}\\
&= \frac{\delta^{-2}(\mathrm{tr}(X_{S\cup v}X_{S\cup v}^T) - \mathrm{tr}(X_S X_S^T))}{(\beta^2 + \delta^{-2}\sigma_1^2(X))^2}\\
&= \frac{\delta^{-2}\mathrm{tr}(X_v X_v^T)}{(\beta^2 + \delta^{-2}\sigma_1^2(X))^2} = \frac{\delta^{-2}}{(\beta^2 + \delta^{-2}\sigma_1^2(X))^2},
\end{aligned}$$

where the third and the fifth equalities are by the definition of the trace of a matrix, the inequality is by Cauchy interlacing inequality of singular values [Strang, 2006], the sixth equality is by the linearity of the trace and the last is by $\mathrm{tr}(X_v X_v^T) = \|X_v\|^2 = 1$. We can similarly derive that

$$f_S(v) \leq \delta^{-2}/\beta^4.$$

According to Definition 3, the lemma thus holds. □

In non-parametric learning (e.g., sparse Gaussian processes), the goal is to select a set of representative data points. Let $C \in \mathbb{R}^{n\times n}$ be the covariance matrix parameterized by a positive definite kernel. Let $C_S \in \mathbb{R}^{|S|\times|S|}$ denote the submatrix of C with its rows and columns indexed by $S \subseteq \{1, 2, \ldots, n\}$. The determinantal function

$$f(S) = \det(I_{|S|} + \delta^{-2}C_S), \quad (13)$$

is often involved in the objective functions of non-parametric learning, e.g., [Lawrence *et al.*, 2003; Kulesza and Taskar, 2012]. Although the logarithm of f is monotone submodular [Krause and Guestrin, 2005], the determinantal function f itself is not submodular in general. Let $\lambda_i(\cdot)$ denote the i-th largest eigenvalue value of a squire matrix. For notational convenience, we will use A and A_S to denote $I_n + \delta^{-2}C$ and $I_{|S|} + \delta^{-2}C_S$, respectively.

Lemma 6. *For the determinantal function (i.e., Eq. (*13*)) in non-parametric learning, the submodularity ratio α can be lower bounded as*

$$\alpha \geq \frac{\lambda_n(A) - 1}{(\lambda_1(A) - 1)\prod_{i=1}^{n-1}\lambda_i(A)}.$$

Proof. For any index set $S \subseteq \{1, \ldots, n\}$ and $v \notin S$, we have

$$\begin{aligned}
f_S(v) &= \det(A_{S\cup v}) - \det(A_S)\\
&= \prod\nolimits_{i=1}^{|S\cup v|}\lambda_i(A_{S\cup v}) - \prod\nolimits_{i=1}^{|S|}\lambda_i(A_S)\\
&\geq (\lambda_{|S\cup v|}(A_{S\cup v}) - 1)\cdot\prod\nolimits_{i=1}^{|S|}\lambda_i(A_S)
\end{aligned}$$

where the inequality is derived by Cauchy interlacing inequality. We can similarly derive that

$$f_S(v) \leq (\lambda_1(A_{S\cup v}) - 1)\cdot\prod\nolimits_{i=1}^{|S|}\lambda_i(A_S).$$

Thus, for any $S \subseteq T$ and $v \notin T$, we have

$$\begin{aligned}
\frac{f_S(v)}{f_T(v)} &\geq \frac{(\lambda_{|S\cup v|}(A_{S\cup v}) - 1)\cdot\prod_{i=1}^{|S|}\lambda_i(A_S)}{(\lambda_1(A_{T\cup v}) - 1)\cdot\prod_{i=1}^{|T|}\lambda_i(A_T)}\\
&\geq \frac{\lambda_{|S\cup v|}(A_{S\cup v}) - 1}{(\lambda_1(A_{T\cup v}) - 1)\cdot\prod_{i=1}^{|T|-|S|}\lambda_i(A_T)}\\
&\geq \frac{\lambda_n(A) - 1}{(\lambda_1(A) - 1)\prod_{i=1}^{n-1}\lambda_i(A)},
\end{aligned}$$

where the last two inequalities are derived by Cauchy interlacing inequality. Thus, the lemma holds. □

7 Conclusion

In this paper, we prove the approximation guarantees of general stochastic greedy algorithms for subset selection with non-monotone or non-submodular objective functions and also with a general constraint. This largely extends previous studies, which mainly focused on subset selection with monotone submodular objective functions and a size constraint. Moreover, we show that the derived approximation guarantees are applicable to real-world subset selection tasks.

Acknowledgments

This work was supported by the NSFC (61603367, 61672478), the Jiangsu NSF (BK20160066), the Science and Technology Innovation Committee Foundation of Shenzhen (ZDSYS201703031748284), and the Royal Society Newton Advanced Fellowship (NA150123).

References

[Bian *et al.*, 2017] A. A. Bian, J. M. Buhmann, A. Krause, and S. Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. In *ICML*, pages 498–507, 2017.

[Buchbinder *et al.*, 2014] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz. Submodular maximization with cardinality constraints. In *SODA*, pages 1433–1452, 2014.

[Calinescu *et al.*, 2011] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

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

[Elenberg *et al.*, 2016] E. R. Elenberg, R. Khanna, A. G. Dimakis, and S. Negahban. Restricted strong convexity implies weak submodularity. *arXiv:1612.00804*, 2016.

[Fisher *et al.*, 1978] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey. An analysis of approximations for maximizing submodular set functions – II. *Polyhedral Combinatorics*, pages 73–87, 1978.

[Hassidim and Singer, 2017] A. Hassidim and Y. Singer. Robust guarantees of stochastic greedy algorithms. In *ICML*, pages 1424–1432, 2017.

[Jenkyns, 1976] T. A. Jenkyns. The efficacy of the 'greedy' algorithm. In *Proceedings of the 7th Southeastern Conference on Combinatorics, Graph Theory and Computing*, pages 341–350, 1976.

[Kempe *et al.*, 2003] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[Khanna *et al.*, 2017] R. Khanna, E. Elenberg, A. Dimakis, S. Negahban, and J. Ghosh. Scalable greedy feature selection via weak submodularity. In *AISTATS*, pages 1560–1568, 2017.

[Korte and Hausmann, 1978] B. Korte and D. Hausmann. An analysis of the greedy heuristic for independence systems. *Annals of Discrete Mathematics*, 2:65–74, 1978.

[Krause and Guestrin, 2005] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *UAI*, pages 324–331, 2005.

[Krause *et al.*, 2008] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

[Kulesza and Taskar, 2012] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

[Lawrence *et al.*, 2003] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In *NIPS*, pages 625–632, 2003.

[Miller, 2002] A. Miller. *Subset Selection in Regression*. Chapman and Hall/CRC, 2nd edition, 2002.

[Mirzasoleiman *et al.*, 2015] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause. Lazier than lazy greedy. In *AAAI*, pages 1812–1818, 2015.

[Nemhauser and Wolsey, 1978] G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research*, 3(3):177–188, 1978.

[Nemhauser *et al.*, 1978] 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.

[Ohsaka and Yoshida, 2015] N. Ohsaka and Y. Yoshida. Monotone k-submodular function maximization with size constraints. In *NIPS*, pages 694–702, 2015.

[Qian *et al.*, 2017a] C. Qian, J.-C. Shi, Y. Yu, and K. Tang. On subset selection with general cost constraints. In *IJCAI*, pages 2613–2619, 2017.

[Qian *et al.*, 2017b] C. Qian, J.-C. Shi, Y. Yu, K. Tang, and Z.-H. Zhou. Subset selection under noise. In *NIPS*, pages 3562–3572, 2017.

[Singla *et al.*, 2016] A. Singla, S. Tschiatschek, and A. Krause. Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. In *AAAI*, pages 2037–2043, 2016.

[Strang, 2006] G. Strang. *Linear Algebra and Its Applications*. Thomson Learning, 4th edition, 2006.

[Zhang and Vorobeychik, 2016] H. Zhang and Y. Vorobeychik. Submodular optimization with routing constraints. In *AAAI*, pages 819–826, 2016.

[Zhou and Spanos, 2016] Y. Zhou and C. Spanos. Causal meets submodular: Subset selection with directed information. In *NIPS*, pages 2649–2657, 2016.