Combinatorial Stochastic-Greedy Bandit

Fares Fourati[1], **Christopher John Quinn**[2], **Mohamed-Slim Alouini**[1], **Vaneet Aggarwal**[3,1]

[1]King Abdullah University of Science and Technology (KAUST)
[2]Iowa State University
[3]Purdue University
fares.fourati@kaust.edu.sa, cjquinn@iastate.edu, slim.alouini@kaust.edu.sa, vaneet@purdue.edu

Abstract

We propose a novel combinatorial stochastic-greedy bandit (SGB) algorithm for combinatorial multi-armed bandit problems when no extra information other than the joint reward of the selected set of n arms at each time step $t \in [T]$ is observed. SGB adopts an optimized stochastic-explore-then-commit approach and is specifically designed for scenarios with a large set of base arms. Unlike existing methods that explore the entire set of unselected base arms during each selection step, our SGB algorithm samples only an optimized proportion of unselected arms and selects actions from this subset. We prove that our algorithm achieves a $(1 - 1/e)$-regret bound of $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{2}{3}} T^{\frac{2}{3}} \log(T)^{\frac{2}{3}})$ for monotone stochastic submodular rewards, which outperforms the state-of-the-art in terms of the cardinality constraint k. Furthermore, we empirically evaluate the performance of our algorithm in the context of online constrained social influence maximization. Our results demonstrate that our proposed approach consistently outperforms the other algorithms, increasing the performance gap as k grows.

Introduction

The stochastic multi-armed bandits (MAB) problem involves selecting an arm in each round t and observing a reward that follows an unknown distribution. The objective is to maximize the accumulated reward within a finite time horizon T. Solving the classical MAB problem requires striking a balance between exploration and exploitation. Should the agent try arms that have been explored less frequently to gather more information (exploration), or should it stick to arms that have yielded higher rewards based on previous observations (exploitation)? An extension of the MAB problem is the combinatorial MAB problem, where, instead of choosing a single arm per round, the agent selects a set of multiple arms and receives a joint reward for that set. When the agent only receives information about the reward associated with the selected set of arms, it is known as *full-bandit feedback* or simply *bandit feedback*. On the other hand, if the agent obtains additional information about the contribution of each arm to the overall reward, it is referred to as *semi-bandit feedback*. The full-bandit feedback setting poses a more significant challenge as the decision-maker has significantly less information than the semi-bandit feedback scenario (Fourati et al. 2023a). This paper focuses on the first scenario for combinatorial MAB, i.e., the bandit feedback setting.

In recent years, there has been growing interest in studying combinatorial multi-armed bandit problems with submodular[1] reward functions (Niazadeh et al. 2020; Nie et al. 2022; Fourati et al. 2023a). The submodularity assumption finds motivation in various real-world scenarios. For instance, opening additional supermarkets in a specific location would result in diminishing returns due to market saturation. As a result, submodular functions are commonly used as objective functions in game theory, economics, and optimization (Fourati et al. 2023a). Submodularity arises in important contexts within combinatorial optimization, such as graph cuts (Goemans and Williamson 1995; Iwata, Fleischer, and Fujishige 2001), rank functions of matroids (Edmonds 2003), and set covering problems (Feige 1998). Some key problems where combinatorial multi-armed bandit problems with submodular reward functions include proposing items with redundant information (Qin and Zhu 2013; Takemori et al. 2020), optimizing client participation in federated learning (Balakrishnan et al. 2022; Fourati et al. 2023b), and social influence maximization without knowledge of the network or diffusion model (Wen et al. 2017; Li et al. 2020; Perrault et al. 2020; Agarwal et al. 2022; Nie et al. 2022).

Similar to previous works (Streeter and Golovin 2008; Golovin, Krause, and Streeter 2014; Niazadeh et al. 2021; Agarwal et al. 2021, 2022; Nie et al. 2022), we assume that the reward function is non-decreasing (monotone) in expectation. Without further constraints, the optimal set will contain all the arms in this setup. Thus, we limit the cardinality of the set to k. Recently, (Nie et al. 2022, 2023) studied this problem and proposed an explore-then-commit greedy (ETCG) algorithm for this problem with full-bandit feedback and showed a $(1 - 1/e)$-regret bound of $\tilde{\mathcal{O}}(n^{\frac{1}{3}} k T^{\frac{2}{3}})$, where n is the number of arms. The algorithm follows a greedy explore-then-commit approach that greedily adds base arms to a super arm (a subset of base arms) until the

[1]A set function $f : 2^{\Omega} \to \mathbb{R}$ defined on a finite set Ω is considered submodular if it exhibits the property of diminishing returns: for any $A \subseteq B \subset \Omega$ and $x \in \Omega \backslash B$, it holds that $f(A \cup x) - f(A) \geq f(B \cup x) - f(B)$.

cardinality constraint is satisfied. It then exploits this super arm for the remaining time. To determine which base arm to add to the super arm, the remaining arms are sampled m times each (where m is a hyper-parameter), and the arm with the highest average reward is chosen. However, in practical scenarios with many arms, exploring all remaining arms in each iteration may require a significant time and thus is unsuitable for smaller T. Therefore, we propose a modified approach where a smaller subset of arms is randomly selected for exploration in each iterative round, and the arm with the highest reward is chosen.

It is worth noting that a similar random selection-based algorithm has been considered in (Mirzasoleiman et al. 2015) for the offline setup, providing a $(1-1/e-\epsilon)$-approximation guarantee, where ϵ determines the reduction in the number of arms selected in each iteration. However, although beneficial for exploration, this sub-selection of arms in each iteration leads to suboptimal approximation guarantees. In this paper, we ask the question: "*Can exploring a subset of arms in each iteration achieve a benefit regarding the* $(1-1/e)$*-regret bound compared to selecting all remaining arms?*"

We answer this question in a positive. By carefully selecting the parameter ϵ, we achieve a $(1-1/e)$-regret bound of $\tilde{\mathcal{O}}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}})$. This improvement in regret bound surpasses that of (Nie et al. 2022, 2023) by orders of magnitude in terms of k. This improvement is particularly significant for larger values of k. Our proposed approach reduces exploration while enhancing expected cumulative $(1-1/e)$-regret performance.

Contributions

We present the main contributions of this paper as follows:

- We introduce stochastic-greedy bandit (SGB), a novel technique in the explore-then-commit greedy strategy with bandit feedback, wherein an optimized proportion of remaining arms are randomly sampled in each greedy iteration. More precisely, rather than sampling $(n-i+1)$ arms in greedy iteration i, random $(n-i+1)\min\{1, \log(1/\epsilon)/k\}$ arms are chosen for an appropriately selected $\epsilon = \mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{-\frac{1}{3}}\log(T)^{-\frac{1}{3}})$, which reduces the amount of exploration.
- We provide theoretical guarantees for SGB by proving that it achieves an expected cumulative $(1-1/e)$-regret of at most $\tilde{\mathcal{O}}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}})$ for monotone stochastic submodular rewards. This represents an improvement of $k^{\frac{1}{3}}$ compared to the previous state-of-the-art method (Nie et al. 2023).
- We conduct empirical experiments to evaluate the performance of our proposed SGB algorithm compared to the previous state-of-the-art algorithms specialized in monotone stochastic submodular rewards (Nie et al. 2022, 2023). We specifically focus on the online social influence maximization problem and demonstrate the efficiency of SGB in achieving superior results in terms of cumulative regret. In particular, the results show that the proposed algorithm outperforms the baselines, with the performance gap increasing as k increases.

Related Work

This section discusses the works closely related to the problem we are investigating. Multi-armed bandits have been considered in two different settings: *adversarial setting*, where an adversary produce a reward succession that may be affected by the agent's prior decisions (Auer et al. 2002), and a *stochastic setting*, where the reward of each action is randomly drawn from a specific distribution, as described in (Auer, Cesa-Bianchi, and Fischer 2002). In this work, we focus on stochastic reward functions. Standard multi-armed bandits find adversarial settings more challenging, and the outcome can be immediately used as one workable method for the stochastic scenario (Lattimore and Szepesvári 2020). However, this is different for submodular bandits. While adversarial environment in adversarial bandits selects a series of submodular functions $\{f_1, \cdots, f_T\}$ (Streeter and Golovin 2008; Golovin, Krause, and Streeter 2014; Roughgarden and Wang 2018; Niazadeh et al. 2020), in the submodular stochastic setting the realizations of the stochastic function in the problem we define need not be submodular, it only needs to be submodular in expectation (Fourati et al. 2023a), meaning the stochastic setting is not a particular case of the adversarial setting.

Submodular maximization has been proven to be NP-hard. Even achieving an approximation ratio of $\alpha \in (1-1/e, 1]$ under a cardinality constraint with access to a monotone submodular value oracle is also NP-hard (Nemhauser, Wolsey, and Fisher 1978; Feige 1998; Feige, Mirrokni, and Vondrák 2011). However, (Nemhauser, Wolsey, and Fisher 1978) proposed a simple greedy $(1-1/e)$-approximation algorithm for monotone submodular maximization under a cardinality constraint. Therefore, the best approximation ratio for monotone submodular objectives with a polynomial time algorithm is $1-1/e$. Thus, we study $(1-1/e)$-regret combinatorial MAB algorithms in this paper.

Table 1 enumerates related combinatorial works with monotone submodular reward function under bandit feedback for both adversarial and stochastic settings. The table summarizes that the proposed approach achieves the state-of-the-art $(1-1/e)$-regret result. Even though we consider stochastic submodular rewards, full-bandit feedback has been studied for non-submodular rewards, including linear reward functions (Dani, Hayes, and Kakade 2008; Rejwan and Mansour 2020) and Lipschitz reward functions (Agarwal et al. 2021, 2022). In these works, the optimal action (best set of k arms) is to use the k individually best arms; that property does not hold for submodular rewards. Further, non-monotone submodular functions with bandit feedback without cardinality constraint have been studied in (Fourati et al. 2023a), where $\frac{1}{2}$-regret is derived. However, this algorithm cannot be directly applied to our setup since it lacks a cardinality constraint.

Recently, (Nie et al. 2023) provided a framework that adapts offline algorithms for combinatorial optimization with a robustness guarantee to online algorithms with provable regret guarantees. We could use this framework for the offline approximation algorithm described in (Mirzasoleiman et al. 2015) for the problem. At the same time, we note that such an approach will result in $(1-1/e-\epsilon)$-

Reference	Stochastic	$(1-1/e)$-Regret
(Streeter and Golovin 2008)		$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^2\, T^{\frac{2}{3}}\,)$
(Golovin, Krause, and Streeter 2014)		$\tilde{\mathcal{O}}(\, n^{\frac{2}{3}}\, k^{\frac{2}{3}}\, T^{\frac{2}{3}}\,)$
(Niazadeh et al. 2021)		$\tilde{\mathcal{O}}(\, n^{\frac{2}{3}}\, k\ \ T^{\frac{2}{3}}\,)$
(Nie et al. 2022)	✓	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{\frac{4}{3}}\, T^{\frac{2}{3}}\,)$
(Nie et al. 2023)	✓	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k\ \ T^{\frac{2}{3}}\,)$
SGB($\epsilon^\star$) (ours)	✓	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{\frac{2}{3}}\, T^{\frac{2}{3}}\,)$

Table 1: Table of selected related works for submodular monotone function maximization under full-bandit feedback. The second column indicates the stochastic or adversarial setting, and the last column gives the expected cumulative $(1-1/e)$-regret bounds. The notation $\tilde{\mathcal{O}}(\cdot)$ drops the $\log$ terms.

approximation because the offline algorithm has $(1-1/e-\epsilon)$ guarantee. Thus, exploring only a subset of arms in each iteration and achieving a $(1-1/e)$-regret is non-trivial and requires careful analysis, which is done in this paper.

Problem Statement

In this section, we present the problem formally. We denote Ω, the ground set of base arms which includes n base arms. We consider decision-making problems with a fixed time horizon T, where the agent, at each time step t, chooses an action $S_t \subseteq \Omega$, with maximum cardinality constraint k. Let $\mathcal{S} = \{S | S \subseteq \Omega \text{ and } |S| \leq k\}$ represent the set of all permitted subsets at any time step.

After deciding the action S_t, the agent acquires reward $f_t(S_t)$. We assume the reward f_t is stochastic, bounded in $[0,1]$, i.i.d. conditioned on a given action, submodular in expectation[2], and monotonically non-decreasing in expectation[3]. The goal of the agent is to maximize the cumulative reward $\sum_{t=1}^{T} f_t(S_t)$. Define the expected reward function as $f(S) = \mathbb{E}[f_t(S)]$, hence $S^\star = \arg\max_{S:|S|\leq k} f(S)$ denote the optimal solution in expectation. One common metric to measure the algorithm's performance is to compare the learner to an agent that has and always chooses the optimal set in expectation $S^\star$.

The best approximation ratio for monotone-constrained submodular objectives with a polynomial time algorithm is $1-1/e$ (Nemhauser, Wolsey, and Fisher 1978). Therefore, we compare the learner's cumulative reward to $(1-1/e)Tf(S^\star)$, and we denote the difference as the $(1-1/e)$-regret, which is defined as follows

$$\mathcal{R}(T) = (1-\frac{1}{e})Tf(S^\star) - \sum_{t=1}^{T} f_t(S_t). \tag{2}$$

[2]A stochastic set function $f: 2^\Omega \to \mathbb{R}$ defined on a finite set Ω is considered submodular in expectation if for all $A \subseteq B \subset \Omega$, and $x \in \Omega \backslash B$, we have,

$$\mathbb{E}[f(A \cup \{x\})] - \mathbb{E}[f(A)] \geq \mathbb{E}[f(B \cup \{x\})] - \mathbb{E}[f(B)]. \tag{1}$$

[3]A stochastic set function $f: 2^\Omega \to \mathbb{R}$ is called non-decreasing in expectation if for any $A \subseteq B \subseteq \Omega$ we have $\mathbb{E}[f(A)] \leq \mathbb{E}[f(B)]$.

Algorithm 1: SGB

Require: ground set Ω, horizon T , cardinality k

$S^{(0)} \leftarrow \emptyset$, $n \leftarrow |\Omega|$

$m \leftarrow \left\lceil \left(\frac{kT}{2n\sqrt{\log(T)}} \right)^{\frac{2}{3}} \right\rceil$, $\epsilon \leftarrow \left(\frac{nk^2}{4T\log(T)} \right)^{\frac{1}{3}}$

$\beta \leftarrow \frac{\log(\frac{1}{\epsilon})}{k}$, $s_1 \leftarrow n \min\{1, \beta\}$

for phase $i \in \{1, \ldots, k\}$ **do**

  $\mathcal{A}_i \leftarrow s_i$ elements sampled from $\Omega \backslash S^{(i-1)}$

  for arm $a \in \mathcal{A}_i$ **do**

    Play $S^{(i-1)} \cup \{a\}$ m times

    Calculate the mean $\bar{f}\left(S^{(i-1)} \cup \{a\}\right)$

  end for

  $a_i \leftarrow \arg\max_{a \in \mathcal{A}_i} \bar{f}\left(S^{(i-1)} \cup \{a\}\right)$

  $S^{(i)} \leftarrow S^{(i-1)} \cup \{a_i\}$

  $s_{i+1} \leftarrow (n-i+1)\min\{1, \beta\}$

end for

for remaining time **do**

  Play $S^{(k)}$

end for

Note that the $(1-1/e)$-regret $\mathcal{R}(T)$ is random and depends on the subsets chosen. In this work, we focus on minimizing the expected cumulative $(1-1/e)$-regret

$$\mathbb{E}[\mathcal{R}(T)] = (1-\frac{1}{e})Tf(S^\star) - \mathbb{E}\left[\sum_{t=1}^{T} f_t(S_t)\right], \tag{3}$$

where the expectation is over both the environment and the sequence of actions.

Proposed Algorithm

This section presents our proposed combinatorial stochastic-greedy bandit (SGB) algorithm that applies our optimized stochastic-explore-then-commit approach. We provide its pseudo-code in Algorithm 1.

The algorithm follows the explore-then-commit structure where base arms are added to a super arm over time greedily until the cardinality constraint is satisfied and then exploits that super arm. However, in contrast to previous explore-then-commit approaches in (Nie et al. 2022, 2023), which

at every exploration phase has a search space of $\mathcal{O}(n)$, to minimize its expected cumulative regret, SGB reduces its search space to $\mathcal{O}(\frac{n}{k}\min\{k, \log(\frac{1}{\epsilon})\})$ arms. The aim of reducing the searched arms in each iteration is to reduce the time spent in the exploration.

Let $S^{(i)}$ represent the super arm when $i < k$ base arms are selected. Our algorithm starts with the empty set, $S^{(0)} = \emptyset$. To add an arm to the set $S^{(i-1)}$, ETCG explores the full subset $\Omega \setminus S^{(i-1)}$. Instead, our procedure explores a smaller subset, i.e., a random subset $\mathcal{A}_i \subseteq \Omega \setminus S^{(i-1)}$. With $\beta = \log(\frac{1}{\epsilon})/k$, the cardinality of $\mathcal{A}_i$ is

$$|\mathcal{A}_i| = s_i = (n - i + 1)\min\{1, \beta\}. \tag{4}$$

For $\beta < 1$, during each exploration phase i, while ETCG explores $(n - i + 1)$ arms, SGB only explores $(n - i + 1)\beta$ arms. Therefore, SGB requires fewer oracle queries per exploration phase than ETCG. For $\beta \geq 1$, during each exploration phase i, SGB explores $(n - i + 1)$ arms, leading it to become a deterministic greedy algorithm and recover the same results as ETCG. We note that $\beta < 1$ happens when $\epsilon > e^{-k}$. Therefore, the lower bound on ϵ exponentially decreases as a function of k, ensuring this is true for most instances. To minimize the cumulative regret, ϵ is optimized as a function of n, k, and T.

Let T_i denote the time step when phase i finishes, for $i \in \{1, \cdots, k\}$. We also denote $T_0 = 0$ and $T_{k+1} = T$ for notational consistency. Let $\bar{f}_t(S)$ denote the empirical mean reward of set S up to and including time t. Let

$$\mathcal{S}_i = \{\ S^{(i-1)} \cup \{a\} :\ a \in \mathcal{A}_i, \mathcal{A}_i \subseteq \Omega \setminus S^{(i-1)}\ \}$$

denote the set of actions considered during phase i. Each action comprises the super arm $S^{(i-1)}$ decided during the last phase and an additional base arm. Each action $S \in \mathcal{S}_i$ is played the same number of times; let m denote that number. The choice of m will be optimized later to minimize regret. At the end of phase $i \in \{1, \ldots, k\}$, SGB selects the action that has the largest empirical mean,

$$a_i = \underset{a \in \mathcal{A}_i}{\arg\max}\ \bar{f}_{T_i}(S^{(i-1)} \cup \{a\}), \tag{5}$$

and include it in the super arm $S^{(i)} = S^{(i-1)} \cup \{a_i\}$. During the final phase, the algorithm exploits $S^{(k)}$; it plays the same action $S_t = S^{(k)}$ for $t \in \{T_k + 1, \cdots, T\}$.

Similar to previous state-of-the-art approaches, SGB has low storage complexity. During exploitation, for $t \in \{T_k + 1, \cdots, T_{k+1}\}$, only the indices of the k base arms are stored, and no additional computation is required. During exploration, for $t \in \{1, \cdots, T_k\}$, for every phase i, SGB needs to store the highest empirical mean and its associated base arm $a \in \mathcal{A}_i$. Therefore, SGB has $\mathcal{O}(k)$ storage complexity. In comparison, the algorithm suggested by (Streeter and Golovin 2008) for the full-bandit adversarial environment has a storage complexity of $\mathcal{O}(nk)$.

We note that the reduction of exploration time through random subset sampling from the remaining arms comes at the expense of reduced offline approximation guarantee to $(1 - 1/e - \epsilon)$ in (Mirzasoleiman et al. 2015). Thus, it is apriori unclear if such an approach can maintain the online $(1 - 1/e)$-regret guarantees with the reduced exploration, which is studied in the next section.

Regret Analysis

This section analyses the regret for Algorithm 1. We begin by stating the main theorem, which bounds the expected cumulative $(1 - 1/e)$-regret of SGB.

Theorem 1. *For the decision-making problem defined in Section 2 with $T \geq n(k+1)\sqrt{\log(T)}$, the expected cumulative $(1 - 1/e)$-regret of SGB is at most $\mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{2}{3}})$.*

The rest of the section provides the proof of this result.

Since for each phase i, we select each action $S^{(i-1)} \cup \{a\} \in \mathcal{S}_i$ exactly m times, we consider the equal-sized confidence radii $\text{rad} = \sqrt{\log(T)/m}$ for all the actions $S^{(i-1)} \cup \{a\} \in \mathcal{S}_i$ at the end of phase i. Denote the event that the empirical means of actions played in phase i are concentrated around their statistical means as

$$\mathcal{E}_i = \bigcap_{S \cup \{a\} \in \mathcal{S}_i} \{|\bar{f}(S \cup \{a\}) - f(S \cup \{a\})| < \text{rad}\}. \tag{6}$$

Then we define the clean event $\mathcal{E}$ to be the event that the empirical means of all actions selected up to and including phase k is within rad of their corresponding statistical means:

$$\mathcal{E} = \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_k.$$

Although the $\mathcal{E}_i$'s are not independent, by conditioning on the sequence of played subsets $\{S^{(0)}, S^{(1)}, \ldots, S^{(k)}\}$ and using the Hoeffding bound (Hoeffding 1994), we show in the Appendix (see (Fourati et al. 2023c)) that $\mathcal{E}$ happens with high probability. We use the concentration of empirical means, Equation (6), and properties of submodularity to show the following result.

Lemma 1. *Under the clean event $\mathcal{E}$, for all $i \in \{1, 2, \cdots, k\}$, for all positive ϵ,*

$$f(S^{(i)}) - f(S^{(i-1)}) \geq \frac{1 - \epsilon}{k}(f(S^\star) - f(S^{(i-1)})) - 2\,\text{rad}\,.$$

Proof. Recall that a_i, defined in (5), is the index of the arm that with $S^{(i-1)}$ forms the action with the highest empirical mean at the end of phase i, and $S^{(i)} = S^{(i-1)} \cup \{a_i\}$. Let a_i^* denote the index of the arm that with $S^{(i-1)}$ forms the action with the highest expected value. For each $a \in \mathcal{A}_i$, the event that the empirical mean $\bar{f}(S^{(i-1)} \cup \{a\})$ is concentrated within a radius of size rad around the expected value. We lower bound the expected reward $f(S^{(i)})$ for the empirically best action in phase i, $S^{(i)} = \{a_i\} \cup S^{(i-1)}$. To do so, we apply (6) to two specific arms, the empirically best a_i out of $\mathcal{A}_i$ and the statistically best $a_i^\star$ out of $\mathcal{A}_i$.

$$\begin{aligned}
f(S^{(i)}) &= f(S^{(i-1)} \cup \{a_i\}) && \text{(by design)} \\
&\geq \bar{f}(S^{(i-1)} \cup \{a_i\}) - \text{rad} && (\text{ using (6)}) \\
&\geq \bar{f}(S^{(i-1)} \cup \{a_i^\star\}) - \text{rad} \\
&\qquad (a_i \text{ has the highest empirical mean}) \\
&\geq f(S^{(i-1)} \cup \{a_i^\star\}) - 2\text{rad}. && (\text{ using (6)})
\end{aligned}$$

Furthermore, using Lemma 2 in (Mirzasoleiman et al. 2015), with $s_i = (n-i+1)\min\{1, \frac{\log(\frac{1}{\epsilon})}{k}\}$, we have

$$f(S^{(i-1)} \cup \{a_i^\star\}) - f(S^{(i-1)}) \geq \frac{1-\epsilon}{k}(f(S^\star) - f(S^{(i-1)}). \tag{7}$$

Combining the above results, we conclude that

$$f(S^{(i)}) - f(S^{(i-1)}) \geq \frac{1-\epsilon}{k}(f(S^\star) - f(S^{(i-1)}) - 2\text{rad}.$$

□

This lemma identifies a lower bound of the expected marginal gain $f\left(S^{(i)}\right) - f\left(S^{(i-1)}\right)$ of the empirically best action $S^{(i)}$ at the end of phase i. As a corollary of Lemma 1, using properties of submodular set functions, we can lower bound the expected value of SGB's chosen set $S^{(k)}$ of size k, which is used for exploitation;

Corollary 1. *Under the clean event $\mathcal{E}$, for all positive ϵ,*

$$f\left(S^{(k)}\right) \geq (1-\frac{1}{e})f(S^\star) - (\epsilon + 2k\text{rad}).$$

Proof. Applying Lemma 1 result recursively for $i = k$, until we get to $S^{(0)} = \emptyset$; $f(\emptyset) = 0$,

$$f(S^{(k)}) \geq \left[\frac{1-\epsilon}{k}f(S^\star) - 2\text{rad}\right]\sum_{\ell=0}^{k-1}(1-\frac{1-\epsilon}{k})^\ell. \tag{8}$$

Simplifying the geometric summation,

$$\sum_{\ell=0}^{k-1}(1-\frac{1}{k})^\ell = \frac{1-(1-\frac{1-\epsilon}{k})^k}{1-(1-\frac{1-\epsilon}{k})} = k\left(1-\left(1-\frac{1-\epsilon}{k}\right)^k\right).$$

Continuing with (8),

$$\begin{aligned} f(S^{(k)}) &\geq \left[\frac{1-\epsilon}{k}f(S^\star) - 2\text{rad}\right] k\left(1-\left(1-\frac{1-\epsilon}{k}\right)^k\right) \\ &\geq \left(1-\left(1-\frac{1-\epsilon}{k}\right)^k\right) f(S^\star) - 2k\text{rad} \\ &\qquad \text{(simplifying with } (1-\tfrac{1}{k})^k \leq 1) \\ &\geq \left(1-e^{-(1-\epsilon)}\right) f(S^\star) - 2k\text{rad}. \end{aligned}$$

Therefore, for $0 \leq \epsilon \leq 1$, using $e^\epsilon \leq 1 + e\epsilon$, we have

$$\begin{aligned} f(S^{(k)}) &\geq (1-\frac{1}{e}-\epsilon)f(S^\star) - 2k\text{rad} \qquad (9) \\ &= (1-\frac{1}{e})f(S^\star) - \epsilon f(S^\star) - 2k\text{rad} \\ &\qquad \text{(rearranging)} \\ &\geq (1-\frac{1}{e})f(S^\star) - \epsilon - 2k\text{rad} \qquad (f(S^\star) \leq 1) \\ &= (1-\frac{1}{e})f(S^\star) - (\epsilon + 2k\text{rad}). \end{aligned}$$

□

We use the above Corollary 1 to bound the expected cumulative regret of our proposed algorithm. We split the expected $(1-1/e)$-regret (3) conditioned on the clean event $\mathcal{E}$ into two parts, one for the exploration and one for the exploitation,

$$\begin{aligned} \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{t=1}^{T}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S_t)]\right) \\ &= \underbrace{\sum_{i=1}^{k}\sum_{t=T_{i-1}+1}^{T_i}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S_t)]\right)}_{\text{Exploration}} \\ &\quad + \underbrace{\sum_{t=T_k+1}^{T}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S^{(k)})]\right)}_{\text{Exploitation}}. \end{aligned} \tag{10}$$

Note that during phase i, each of the s_i actions in $\mathcal{S}_i$ is selected exactly m times, thus $T_i - T_{i-1} = ms_i$. For each action S_t choosed during phase i, that is for $t \in \{T_{i-1}+1, \cdots, T_i\}$, since $S^{(i-1)} \subset S_t$, by monotonicity of the expected reward function f we have $f(S^{(i-1)}) \leq f(S_t)$. Thus we can upper bound the expected regret $\mathbb{E}[\mathcal{R}(T)|\mathcal{E}]$ incurred during the first k phases (first term of (10)) as

$$\begin{aligned} &\sum_{i=1}^{k}\sum_{t=T_{i-1}+1}^{T_i}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S_t)]\right) \\ &\leq \sum_{i=1}^{k} ms_i\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S^{(i-1)})]\right) \\ &\leq ms_1\sum_{i=1}^{k}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S^{(i-1)})]\right) \\ &\leq ms_1 k. \end{aligned} \tag{11}$$

The last inequality follows because the rewards are in $[0,1]$.

We can upper bound the expected regret $\mathbb{E}[\mathcal{R}(T)|\mathcal{E}]$ incurred during the exploitation phase (phase $k+1$; second term of (10)) by applying Corollary 1 as follows

$$\begin{aligned} &\sum_{t=T_k+1}^{T}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S^{(k)})]\right) \\ &\leq \sum_{t=T_k+1}^{T}(\epsilon + 2k\text{rad}) \\ &\leq T\epsilon + 2kT\text{rad}. \end{aligned} \tag{12}$$

Combining the upper bounds (11) and (12), we get

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq ms_1 k + \epsilon T + 2kT\text{rad}. \tag{13}$$

We have $s_1 = n\min\{1, \frac{\log(\frac{1}{\epsilon})}{k}\}$ and $\text{rad} = \sqrt{\frac{\log(T)}{m}}$. Therefore, we have

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq mn\min\{k, \log(\frac{1}{\epsilon})\} + \epsilon T + 2kT\sqrt{\frac{\log(T)}{m}}. \tag{14}$$

First, trivially $\min\{k, \log(\frac{1}{\epsilon})\} \leq \log(\frac{1}{\epsilon})$, thus

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq mn\log(\frac{1}{\epsilon}) + \epsilon T + 2kT\sqrt{\frac{\log(T)}{m}}. \quad (15)$$

Setting the derivative (with respect to ϵ) of the bound to 0,

$$0 = -mn\frac{1}{\epsilon} + T + 0 \qquad \Rightarrow \qquad \epsilon = \frac{mn}{T}. \quad (16)$$

The second derivative (with respect to ϵ), $mn\epsilon^{-2}$, is positive so the stationary point is a minimizer.

Plugging $\epsilon = \frac{mn}{T}$ into the regret upper bound,

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq mn\log(\frac{T}{mn}) + \frac{mn}{T}T + 2kT\sqrt{\frac{\log(T)}{m}} \\ &\leq mn\log(T) + mn + 2kT\sqrt{\frac{\log(T)}{m}} \\ &\leq 2mn\log(T) + 2kT\sqrt{\frac{\log(T)}{m}}. \quad (17)\end{aligned}$$

The above inequality is valid for all m strictly greater than zero. Hence, to find a tighter bound, we find $m^\star$ that minimizes the right side. Thus we get

$$m^\star = \left(\frac{kT\sqrt{\log(T)}}{2n\log(T)}\right)^{\frac{2}{3}} = \left(\frac{kT}{2n\sqrt{\log(T)}}\right)^{\frac{2}{3}}. \quad (18)$$

For $T \geq n(k+1)\sqrt{\log(T)}$, we have $m^\star \geq \frac{1}{2}$, therefore $\lceil m^\star \rceil \leq 2m^\star$. Plugging $m = \lceil m^\star \rceil$ into the regret bound,

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq \lceil m^\star \rceil n\log(T) + 2kT\log(T)^{1/2}\lceil m^\star \rceil^{-1/2} \\ &\leq 2m^\star n\log(T) + 2kT\log(T)^{1/2}(m^\star)^{-1/2} \\ &= 2^{\frac{1}{3}}k^{\frac{2}{3}}T^{2/3}n^{-2/3}\log(T)^{-1/3}n\log(T) \\ &+ 2^{\frac{4}{3}}k^{\frac{2}{3}}T\log(T)^{1/2}T^{-1/3}n^{1/3}\log(T)^{1/6} \\ &\leq \mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{2}{3}}). \quad (19)\end{aligned}$$

Based on $m^\star$, we define the optimal $\epsilon^\star$ as follows

$$\epsilon^\star = \frac{m^\star n}{T} = nT^{-1}\left(\frac{kT}{2n\sqrt{\log(T)}}\right)^{\frac{2}{3}} = \left(\frac{nk^2}{4T\log(T)}\right)^{\frac{1}{3}}.$$

Under the bad event, i.e., the complement $\bar{\mathcal{E}}$ of the good event $\mathcal{E}$, given that the rewards are bounded in $[0,1]$, it can be easily seen that $\mathbb{E}[\mathcal{R}(T) \mid \bar{\mathcal{E}}] \leq T$. Moreover, by using the Hoeffding inequality (Hoeffding 1994), for $T \geq nk$, we have $\mathbb{P}(\bar{\mathcal{E}}) \leq \frac{2}{T}$, see Appendix (available at (Fourati et al. 2023c)). Therefore, we obtain $\mathbb{E}[\mathcal{R}(T)] \leq \mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{2}{3}})$.

Remark 1. *The framework of Nie et al. (2023) that adapts offline algorithms for combinatorial optimization problems with robustness guarantees to online settings via the explore-then-commit approach can be applied to the offline algorithm in (Mirzasoleiman et al. 2015). However, as this offline algorithm has $(1-1/e-\epsilon)$-approximation guarantee, such an approach will give a weaker $(1-1/e-\epsilon)$-regret guarantee rather than $(1-1/e)$-regret guarantee studied in this paper.*

Remark 2. *For unknown time horizon T, the geometric doubling trick can extend our result to an anytime algorithm. To initialize the algorithm, we choose T_0 to be large enough, then we choose a geometric succession $T_i = T_0 2^i$ for $i \in \{1, 2, \cdots\}$, and run our algorithm during the time interval $T_{i+1} - T_i$ with a complete restart. From Theorem 4 in (Besson and Kaufmann 2018), we can prove that the regret bound preserves the $T^{2/3}$ dependency with changes only in the constant factor.*

Remark 3. *For the scenario we study in this paper of combinatorial multi-armed bandit with submodular rewards in expectation and under full-bandit feedback, it is still unknown if $\tilde{\mathcal{O}}(T^{1/2})$ expected cumulative $(1-1/e)$-regret is possible (ignoring n and k dependence), and only $\tilde{\mathcal{O}}(T^{2/3})$ bounds have been shown in the literature; see Table 1.*

Experiments on Online Social Influence Maximization

Problem Statement

Social influence maximization is a combinatorial problem, which consists of selecting a subset of nodes in a graph that can influence the remaining nodes. For instance, when marketing a newly developed product, one strategy is to identify a group of highly influential individuals and rely on their recommendations to reach a broader audience. Influence maximization can be formulated as a monotone submodular maximization problem, where adding more nodes to the selected set yields diminishing returns without negatively affecting other nodes. Typically, there is a fixed constraint on the cardinality of the selected set. While some works have addressed influence maximization as a multi-armed bandit problem with additional feedback (Lei et al. 2015; Wen et al. 2017; Vaswani et al. 2017; Li et al. 2020; Perrault et al. 2020), this feedback is often unavailable in most social networks, except for a few public accounts. Recently, (Nie et al. 2022) proposed the ETCG algorithm for influence maximization under full-bandit feedback. Their algorithm demonstrated superior performance through empirical evaluations compared to other full-bandit algorithms. In this work, we compare our SGB method, for different ϵ values, including the optimized value $\epsilon^\star = (\frac{nk^2}{4T\log(T)})^{\frac{1}{3}}$, with ETCG (Nie et al. 2022, 2023).

Experiment Details

For the experiments, instead of $(1-1/e)$ regret in Eq. (2), which requires knowing $S^\star$, we compare the cumulative rewards achieved by SGB for different ϵ, including $\epsilon^\star$, and ETCG against $Tf(S^{\mathrm{grd}})$, where S^{grd} is the solution returned by the offline $(1-1/e)$-approximation algorithm suggested by (Nemhauser, Wolsey, and Fisher 1978). Since $f(S^{\mathrm{grd}}) \geq (1-1/e)f(S^\star)$, thus $Tf(S^{\mathrm{grd}})$ is a more challenging reference value than $(1-1/e)Tf(S^\star)$.

We experimented using a portion of the Facebook network (Leskovec and Mcauley 2012). We used the community detection method proposed by (Blondel et al. 2008) to detect a community with 534 nodes and 8158 edges, enabling multiple experiments for various horizons. The dif-

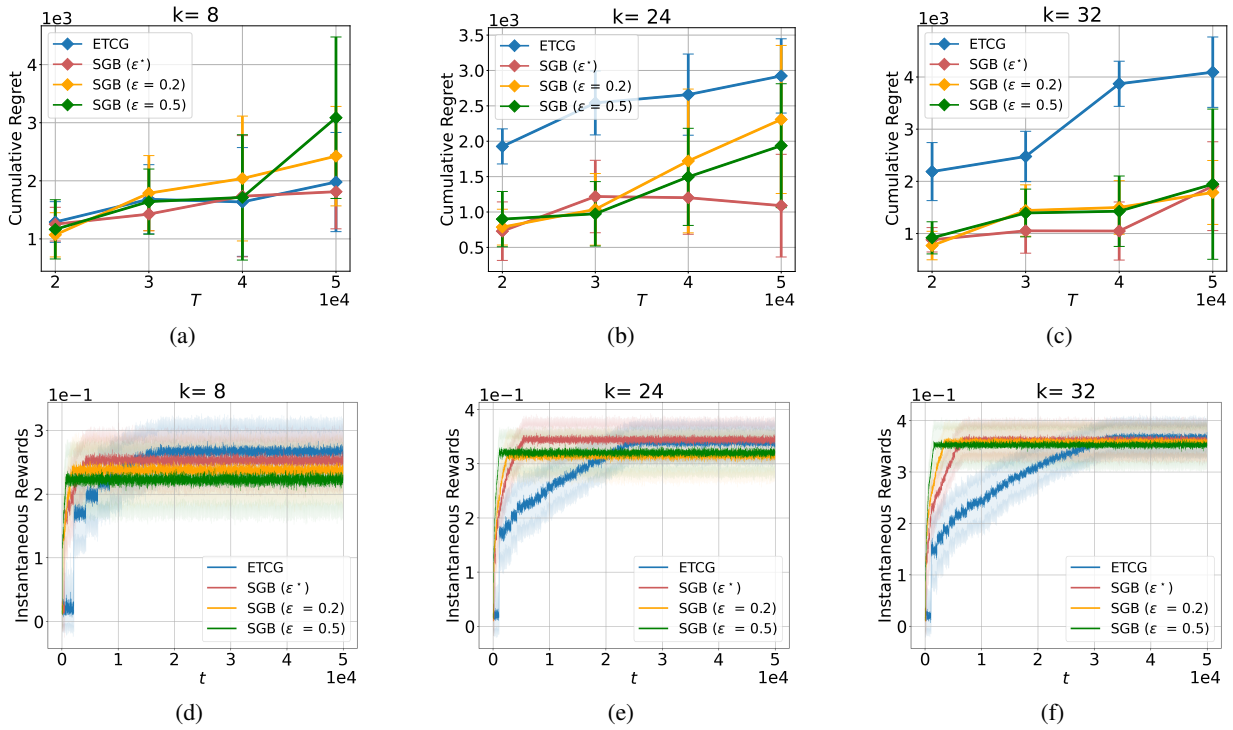

Figure 1: Comparison of SGB, for different ϵ values, including $\epsilon^\star = (\frac{nk^2}{4T\log(T)})^{\frac{1}{3}}$, and ETCG. (a), (b), and (c) are the cumulative regret results as a function of horizon T. (d), (e), and (f) show the moving average plots of the immediate rewards as a function of t, with a window size of 100, with T fixed at 5×10^4, for which the respective $\epsilon^\star$ values are around 0.251, 0.522, and 0.632.

fusion process is simulated using the independent cascade model (Kempe, Kleinberg, and Tardos 2003), wherein in each discrete step, an active node (that was inactive at the previous time step) independently tries to infect each of its inactive neighbors. We used 0.1 uniform infection probabilities for each edge. For every time horizon $T \in \{2\times10^4, 3\times 10^4, 4\times10^4, 5\times10^4\}$, we tested each method ten times.

Experimental Results

Figures (1a), (1b), and (1c) show average cumulative regret curves for SGB with different values of the parameter ϵ, including the optimal value $\epsilon^\star = (\frac{nk^2}{4T\log(T)})^{\frac{1}{3}}$, for various time horizon values T, with a cardinality constraint k set to 8, 24, and 32, respectively. The shaded areas depict the standard deviation. The figure axes are linearly scaled, so a linear cumulative regret curve corresponds to a linear $\widetilde{O}(T)$ cumulative regret. When $k = 8$, SGB with $\epsilon^\star$ demonstrates nearly the lowest average cumulative regret across different time horizons T. However, with non-optimal values of ϵ (0.2 and 0.5), the cumulative regret of SGB is higher than that of ETCG. For higher values of k, such as 24 and 32, with $\epsilon^\star$ as shown in Figures (1b) and (1c), respectively, SGB with all the considered ϵ values outperforms ETCG with lower average cumulative regrets. Furthermore, Figures (1d), (1e), and (1f) illustrate immediate rewards over a horizon $T = 5\times10^4$ for cardinality constraints k of 8, 24, and 32, and $\epsilon^\star$ values around 0.251, 0.522, and 0.632, respectively. The curves for all methods are smoothed using a moving average with a window size of 100. For $k = 8$, although ETCG finds a slightly better solution, SGB with all ϵ ends exploration much faster. For $k = 24$, as shown in Fig. (1e), SGB using $\epsilon^\star$ ends exploration much faster than ETCG and achieves a better solution. Using other ϵ values ends exploration slightly faster than the optimal value but to a lower solution. Similarly, for $k = 32$, as shown in Fig. (1f), SGB with different ϵ values ends exploration 30 times faster than ETCG to a solution within a 0.01 neighborhood of 0.37. Furthermore, using $\epsilon^\star$ yields the best result compared to other values. Therefore, as predicted by the theory, SGB using $\epsilon^\star$ has lower expected cumulative regret than ETCG. Additionally, as observed in the experiments and predicted by the theory, our method becomes more effective for larger values of k.

Conclusion

This paper introduces SGB, a novel method in the online greedy strategy, which incorporates subset random sampling from the remaining arms in each greedy iteration. Theoretical analysis establishes that SGB achieves an expected cumulative $(1-1/e)$-regret of at most $\tilde{\mathcal{O}}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}})$ for monotone stochastic submodular rewards, outperforming the previous state-of-the-art method by a factor of $k^{1/3}$ (Nie et al. 2023). Empirical experiments on online influence maxi-

mization shows SGB's superior performance, highlighting its effectiveness and potential for real-world applications.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation under Grants CCF-2149588 and CCF-2149617 and by Cisco Inc.

References

Agarwal, M.; Aggarwal, V.; Umrawal, A. K.; and Quinn, C. 2021. DART: Adaptive Accept Reject Algorithm for Non-Linear Combinatorial Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8): 6557–6565.

Agarwal, M.; Aggarwal, V.; Umrawal, A. K.; and Quinn, C. J. 2022. Stochastic Top K-Subset Bandits with Linear Space and Non-Linear Feedback with Applications to Social Influence Maximization. *ACM/IMS Transactions on Data Science (TDS)*, 2(4): 1–39.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2): 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77.

Balakrishnan, R.; Li, T.; Zhou, T.; Himayat, N.; Smith, V.; and Bilmes, J. 2022. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*.

Besson, L.; and Kaufmann, E. 2018. What Doubling Tricks Can and Can't Do for Multi-Armed Bandits. *ArXiv*, abs/1803.06971.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008: 10008.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, 355–366.

Edmonds, J. 2003. Submodular functions, matroids, and certain polyhedra. In *Combinatorial Optimization—Eureka, You Shrink!*, 11–26. Springer.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4): 634–652.

Feige, U.; Mirrokni, V. S.; and Vondrák, J. 2011. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4): 1133–1153.

Fourati, F.; Aggarwal, V.; Quinn, C.; and Alouini, M.-S. 2023a. Randomized greedy learning for non-monotone stochastic submodular maximization under full-bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, 7455–7471. PMLR.

Fourati, F.; Kharrat, S.; Aggarwal, V.; Alouini, M.-S.; and Canini, M. 2023b. FilFL: Accelerating Federated Learning via Client Filtering. *arXiv preprint arXiv:2302.06599*.

Fourati, F.; Quinn, C. J.; Alouini, M.-S.; and Aggarwal, V. 2023c. Combinatorial Stochastic-Greedy Bandit. *arXiv preprint arXiv:2312.08057*.

Goemans, M. X.; and Williamson, D. P. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6): 1115–1145.

Golovin, D.; Krause, A.; and Streeter, M. 2014. Online submodular maximization under a matroid constraint with application to learning assignments. *arXiv preprint arXiv:1407.1082*.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, 409–426. Springer.

Iwata, S.; Fleischer, L.; and Fujishige, S. 2001. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM (JACM)*, 48(4): 761–777.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Lei, S.; Maniu, S.; Mo, L.; Cheng, R.; and Senellart, P. 2015. Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 645–654.

Leskovec, J.; and Mcauley, J. 2012. Learning to Discover Social Circles in Ego Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Li, S.; Kong, F.; Tang, K.; Li, Q.; and Chen, W. 2020. Online Influence Maximization under Linear Threshold Model. *arXiv preprint arXiv:2011.06378*.

Mirzasoleiman, B.; Badanidiyuru, A.; Karbasi, A.; Vondrák, J.; and Krause, A. 2015. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming*, 14(1): 265–294.

Niazadeh, R.; Golrezaei, N.; Wang, J.; Susan, F.; and Badanidiyuru, A. 2020. Online Learning via Offline Greedy: Applications in Market Design and Optimization. *EC 2021, Management Science Journal*.

Niazadeh, R.; Golrezaei, N.; Wang, J. R.; Susan, F.; and Badanidiyuru, A. 2021. Online learning via offline greedy algorithms: Applications in market design and optimization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, 737–738.

Nie, G.; Agarwal, M.; Umrawal, A. K.; Aggarwal, V.; and Quinn, C. J. 2022. An Explore-then-Commit Algorithm for Submodular Maximization Under Full-bandit Feedback. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Nie, G.; Nadew, Y. Y.; Zhu, Y.; Aggarwal, V.; and Quinn, C. J. 2023. A Framework for Adapting Offline Algorithms to Solve Combinatorial Multi-Armed Bandit Problems with Bandit Feedback. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.

Perrault, P.; Healey, J.; Wen, Z.; and Valko, M. 2020. Budgeted online influence maximization. In *International Conference on Machine Learning*, 7620–7631. PMLR.

Qin, L.; and Zhu, X. 2013. Promoting Diversity in Recommendation by Entropy Regularizer. In *IJCAI*.

Rejwan, I.; and Mansour, Y. 2020. Top-k Combinatorial Bandits with Full-Bandit Feedback. In *Algorithmic Learning Theory*, 752–776.

Roughgarden, T.; and Wang, J. R. 2018. An Optimal Learning Algorithm for Online Unconstrained Submodular Maximization. In *Proceedings of the 31st Conference On Learning Theory*, 1307–1325.

Streeter, M.; and Golovin, D. 2008. An Online Algorithm for Maximizing Submodular Functions. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, 1577–1584. Red Hook, NY, USA: Curran Associates Inc.

Takemori, S.; Sato, M.; Sonoda, T.; Singh, J.; and Ohkuma, T. 2020. Submodular Bandit Problem Under Multiple Constraints. In *Conference on Uncertainty in Artificial Intelligence*, 191–200. PMLR.

Vaswani, S.; Kveton, B.; Wen, Z.; Ghavamzadeh, M.; Lakshmanan, L. V.; and Schmidt, M. 2017. Model-independent online learning for influence maximization. In *International Conference on Machine Learning*, 3530–3539. PMLR.

Wen, Z.; Kveton, B.; Valko, M.; and Vaswani, S. 2017. Online influence maximization under independent cascade model with semi-bandit feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3026–3036.