Combinatorial Stochastic-Greedy Bandit

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a combinatorial stochastic-greedy bandit (SGB) algorithm for combinatorial multi-armed bandit problems, where only bandit feedback is available. The algorithm is designed explicitly for settings with a large set of base arms in which current methods perform poorly unless the horizon T is large. SGB applies our proposed optimized-stochastic-explore-then-commit approach. Instead of exploring the whole set of base arms, the SGB algorithm iteratively randomly samples only an optimized proportion of arms and then chooses actions from them. Compared to standard methods, SGB reduces its search space from $\mathcal{O}(n)$ choices to $\mathcal{O}(\frac{n}{k}\min\{k, \log(\frac{T}{nk^2})\})$ choices, where n is the number of arms and k is the cardinality constraint. Therefore, by design, SGB has a significantly faster exploration phase, resulting in lower expected cumulative regret. We demonstrate that SGB has a $(1-\frac{1}{e})$-regret bound with an order of $\mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{1}{3}})$ for stochastic submodular rewards, which outperforms the state-of-the-art in terms of the parameter k. Furthermore, we empirically demonstrate the effectiveness of our algorithm in the context of online constrained social influence maximization.

1 Introduction

The stochastic multi-armed bandits, initially proposed by Robbins [1952], model a variety of complex decision-making issues, including clinical choices, investing, setting prices, influence maximization, and product recommendation. A decision maker's objective may be maximizing a specific reward function that depends on her actions over time [Thompson, 1933, Auer, 2002, Auer et al., 2002a, Auer and Ortner, 2010, Agrawal and Goyal, 2012, Gopalan et al., 2014]. The agent must explore several options to reduce the risk of sticking to a sub-optimal decision. Exploration implies, in some ways, making suboptimal decisions over time, which may be costly, resulting in increased regret of not picking the ideal arm, but that is unavoidable to identify the optimal action. As a result, the agent is forced to balance exploration and exploitation. Several problems arise, including how long the agent should explore, how to investigate, and what guarantees we can obtain.

The combinatorial multi-armed bandit problem is a natural extension of the multi-armed bandit problem. Instead of choosing just one individual arm per round, the agent selects a set of multiple arms and receives a joint reward for that set. If the only information that the agent receives is the reward for the selected set of arms, this is referred to as *full-bandit feedback* or simply *bandit feedback*. Otherwise, if the agent receives additional information about their decision, such as the contribution of each arm in the selected set to the overall reward, it is called *semi-bandit feedback*. The former setting is more challenging because the decision-maker has significantly less information than in the latter scenario [Fourati et al., 2023a]. This study focuses on the first scenario, i.e., under bandit feedback.

Recently, there has been much interest in studying combinatorial multi-armed bandit problems with submodular reward functions [Niazadeh et al., 2020, Nie et al., 2022, Fourati et al., 2023a]. A set

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

function $f : 2^{\Omega} \to \mathbb{R}$ defined on a finite set Ω is considered submodular if it has the diminishing returns property: for any $A \subseteq B \subset \Omega$ and $x \in \Omega \backslash B$, it holds that $f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$. Several real-world events motivate the submodularity assumption. For example, opening more supermarkets in a specific location would result in diminishing returns owing to demand saturation. As a result, submodular functions are often used as objective functions in game theory and economics. Recently, submodular maximization algorithms were used for optimizing client participation in federated learning through client filtering [Fourati et al., 2023b] and client selection [Balakrishnan et al., 2022]. Moreover, submodularity emerges in a variety of significant situations in combinatorial optimization, including graph cuts [Goemans and Williamson, 1995, Iwata et al., 2001], rank functions of matroids [Edmonds, 2003], and set covering problems [Feige, 1998]. Furthermore, proposing items with redundant information results in decreased utility returns; hence, this topic has been studied using submodular combinatorial multi-armed bandits [Qin and Zhu, 2013, Yue and Guestrin, 2011, Takemori et al., 2020]. A recent line of research has looked at the challenge of social influence maximization as a multi-armed bandit problem that does not require knowledge of the network or diffusion model [Lei et al., 2015, Wen et al., 2017, Vaswani et al., 2017, Li et al., 2020, Perrault et al., 2020].

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., 2002b], and a *stochastic setting*, where the reward of each action is randomly drawn from a specific distribution, as described in [Auer et al., 2002a]. In this work, we focus on stochastic reward functions. Thus, similar to [Nie et al., 2022, Fourati et al., 2023a], we assume a more relaxed property than submodularity, which is submodularity in expectation, defined as follows.

Definition 1. *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,*

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

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 $\{f_1, ..., f_T\}$ [Streeter and Golovin, 2008, Golovin et al., 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, meaning the stochastic setting is not a special case of the adversarial setting.

Furthermore, similar to previous works [Streeter and Golovin, 2008, Golovin et al., 2014, Niazadeh et al., 2021, Agarwal et al., 2021b,a, Nie et al., 2022], we assume that the reward function is non-decreasing (monotone) in expectation. More precisely, 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 $\mathbb{E}[f(A)] \leq \mathbb{E}[f(B)]$. In contrast to the (non-monotone) work of Fourati et al. [2023a], which finds a subset of arms of any size (unconstrained), this work considers finding a subset of arms with at most k arms (constrained).

1.1 Contributions

We summarize the main contributions of this paper as follows:

- We introduce a new technique that optimizes the bandit search space based on problem-specific parameters such as n, k, and T, to minimize the expected cumulative regret.
- We propose a combinatorial stochastic-greedy bandit (SGB), under bandit feedback, which applies our optimized-stochastic-explore-then-commit approach which outperforms the explore-then-commit approach.
- We prove that SGB guarantees an expected cumulative $(1 - \frac{1}{e})$-regret of at most $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{2}{3}} T^{\frac{2}{3}} \log(T)^{\frac{1}{3}})$ for monotone stochastic submodular rewards. This represents an improvement of $k^{1/3}$ compared to the previous state-of-the-art method [Nie et al., 2023].
- We empirically demonstrate the efficiency of our proposed SGB algorithm against previous state-of-the-art algorithm specialized in monotone stochastic submodular rewards [Nie et al., 2022, 2023] in the case of online social influence maximization.

Combinatorial Bandit Work	Objective			Analysis	
	Submodular	Constrained	Stochastic	Search	α-Regret Bound
Streeter and Golovin [2008]	✓	✓		$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{2}\, T^{\frac{2}{3}}\,)$
Golovin et al. [2014]	✓	✓		$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{2}{3}}\, k^{\frac{2}{3}}\, T^{\frac{2}{3}}\,)$
Niazadeh et al. [2021]	✓	✓		$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{2}{3}}\, k\, T^{\frac{2}{3}}\,)$
Agarwal et al. [2021b]		✓	✓	$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{2}}\, k^{\frac{3}{2}}\, T^{\frac{1}{2}}\,)$
Agarwal et al. [2021a]		✓	✓	$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{\frac{1}{2}}\, T^{\frac{2}{3}}\,)$
Nie et al. [2022]	✓	✓	✓	$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{\frac{4}{3}}\, T^{\frac{2}{3}}\,)$
Fourati et al. [2023a]	✓		✓	$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n\, T^{\frac{2}{3}}\,)$
Nie et al. [2023]	✓	✓	✓	$\tilde{\mathcal{O}}(n)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k\, T^{\frac{2}{3}}\,)$
SGB($\varepsilon^\star$) (ours)	✓	✓	✓	$\tilde{\mathcal{O}}(n/k)$	$\tilde{\mathcal{O}}(\, n^{\frac{1}{3}}\, k^{\frac{2}{3}}\, T^{\frac{2}{3}}\,)$

Table 1: Table of selected related works for combinatorial maximization under full-bandit feedback, enumerating their different assumptions, respective search space of every exploration round, and respective regret guarantees. For all the works above, the approximation ratio, $\alpha = 1 - 1/e$, except for the unconstrained work of Fourati et al. [2023a] where $\alpha = 1/2$. The notation $\tilde{\mathcal{O}}(\cdot)$ drops the $\log$ terms.

1.2 Related Work

This section discusses the works closely related to the problem we are investigating. Our work focuses on submodular, constrained, and stochastic objectives. Table 1 enumerates related combinatorial works under bandit feedback with their respective reward characteristics. We provide the search space for every exploration round and the regret upper bounds for the different methods, as a function of n, k, and T, with n the number of base arms, k the cardinality constraint, and T the time horizon.

Submodular maximization has been proven to be NP-hard. Even achieving an approximation ratio of $\alpha \in (1-\frac{1}{e}, 1]$ under a cardinality constraint with access to a monotone submodular value oracle is also NP-hard [Nemhauser et al., 1978, Feige, 1998, Feige et al., 2011]. However, [Nemhauser et al., 1978] proposed a simple greedy $(1 - \frac{1}{e})$-approximation algorithm for non-negative monotone submodular maximization under a cardinality constraint. Recently, [Nie et al., 2022] extended [Nemhauser et al., 1978]'s offline algorithm with $(1 - \frac{1}{e})$-approximation guarantees to the online setting under full-bandit feedback with their explore-then-commit greedy (ETCG) algorithm for monotone stochastic submodular rewards. They proved $(1 - \frac{1}{e})$-regret guarantees of $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{4}{3}} T^{\frac{2}{3}} \log(T)^{\frac{1}{2}})$. Similarly, [Fourati et al., 2023a] proposed an online explore-then-commit approach that adapts the offline algorithm of [Buchbinder et al., 2015] with $\frac{1}{2}$-approximation guarantees for unconstrained stochastic submodular rewards and proved a $\frac{1}{2}$-regret upper bound of $\mathcal{O}(nT^{\frac{2}{3}} \log(T)^{\frac{1}{3}})$. More recently, [Nie et al., 2023] introduced a general framework for adapting offline algorithms with α-approximation gurantees to the online setting with α-regret guarantees using the same explore-then-commit approach. For monotone submodular rewards, [Nie et al., 2023] tightened the regret of ETCG to $(1 - \frac{1}{e})$-regret guarantees of $\mathcal{O}(n^{\frac{1}{3}} k T^{\frac{2}{3}} \log(T)^{\frac{1}{2}})$.

In our work, we propose a new optimized-stochastic-explore-then-commit approach and use it to adapt the offline algorithm of [Mirzasoleiman et al., 2015] to the online setting. Our proposed online algorithm achieves state-of-the-art results with $(1 - \frac{1}{e})$-regret guarantees of $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{2}{3}} T^{\frac{2}{3}} \log(T)^{\frac{1}{2}})$ for monotone constrained stochastic submodular maximization. While the explore-then-commit approach extends offline algorithms with α-approximation guarantees to α-regret guarantees. Our optimized-stochastic-explore-then-commit approach noticeably extends the offline algorithm of [Mirzasoleiman et al., 2015] with $(\alpha - \varepsilon)$-approximation guarantees to α-regret guarantees for $\alpha = 1 - \frac{1}{e}$ and $\varepsilon = (\frac{nk^2}{4T \log(T)})^{1/3}$, which is found by optimizing the α-regret over the parameter ε.

The framework proposed in [Nie et al., 2023] can be utilized as well to adapt the offline algorithm described in [Mirzasoleiman et al., 2015] for online settings using the explore-then-commit approach. However, since the offline algorithm already guarantees a $(1 - \frac{1}{e} - \varepsilon)$-approximation, applying the framework would only yield a $(1 - \frac{1}{e} - \varepsilon)$-regret guarantee, which is weaker than the $(1 - \frac{1}{e})$-regret guarantee provided by our approach.

2 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 [1] $S_t \subseteq \Omega$, 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 and monotonically non-decreasing in expectation. The goal of the agent is to maximise 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 possible approximation ratio for monotone constrained submodular objectives with a polynomial time algorithm is $1 - \frac{1}{e}$ [Nemhauser et al., 1978]. Therefore, we compare the learner's cumulative reward to $(1 - \frac{1}{e})Tf(S^\star)$, and we denote the difference as the $(1 - \frac{1}{e})$-regret, which is defined as follow

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

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

$$\mathbb{E}[\mathcal{R}(T)] = (1 - \frac{1}{e})Tf(S^\star) - \mathbb{E}\left[\sum_{t=1}^{T} f_t(S_t)\right], \tag{2}$$

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

3 Proposed Algorithm

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

Like explore-then-commit bandits [Nie et al., 2023], such as ETCG [Nie et al., 2022] and RGL [Fourati et al., 2023a], SGB has two phases, an exploration phase and an exploitation one. Similar to the cardinality-constrained algorithm ETCG [Nie et al., 2022, 2023], during exploration, while the cardinality constraint is not yet satisfied, our algorithm adds base arms to a subset of base arms (super arm) greedily over time, then exploits it until horizon T.

In contrast to previous explore-then-commit approaches, which at every exploration 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.

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 $\mathcal{A}_i \subseteq \Omega \setminus S^{(i-1)}$.

Algorithm 1 SGB

Require: ground set Ω, horizon T , cardinality k
  $S^{(0)} \leftarrow \emptyset$, $n \leftarrow |\Omega|$
  $m \leftarrow \left\lceil \left(\frac{kT}{2n\sqrt{\log(T)}}\right)^{2/3} \right\rceil$, $\varepsilon \leftarrow \left(\frac{nk^2}{4T\log(T)}\right)^{1/3}$
  $\beta \leftarrow \frac{\log(\frac{1}{\varepsilon})}{k}$, $s_1 \leftarrow n\min\{1, \beta\}$
  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

[1] We use the *super-arm*, *subset* and *action* terminologies interchangeably throughout the paper.

With $\beta = \log(\frac{1}{\varepsilon})/k$, the cardinality of $\mathcal{A}_i$ is

$$|\mathcal{A}_i| = s_i = (n - i + 1)\min\{1, \beta\}. \tag{3}$$

For $\beta < 1$, during each exploration phase i, while ETCG explores $n - i + 1$ arms, SGB only explores $(n-i+1)\beta$ arms. This means that SGB requires fewer oracle queries per exploration phase compared to 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. It is worth noting that in most settings, $\beta < 1$ is more common. In fact, $\beta < 1$ if and only if $\varepsilon > \exp(-k)$. Therefore, the lower bound on ε exponentially decreases as a function of k, ensuring that this is true for most instances. Choosing a random set $\mathcal{A}_i$ of size s_i is sufficient to overlap with the optimal set $S^\star$ with probability $1 - \varepsilon$ [Mirzasoleiman et al., 2015]. The optimal ε should depend on the setting (problem); thus, we optimize it as a function of n, k, and T to minimize the $(1 - \frac{1}{e})$-regret.

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,

$$a_i = \operatorname*{arg\,max}_{a \in \mathcal{A}_i}\ \bar{f}_{T_i}(S^{(i-1)} \cup \{a\}), \tag{4}$$

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 $a \in \mathcal{A}_i \setminus S^{(i)}$. 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 $\mathcal{O}(nk)$.

4 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 $T \geq nk\sqrt{\log(T)/2}$, the expected cumulative $(1 - 1/e)$-regret of SGB is at most $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{2}{3}} T^{\frac{2}{3}} \log(T)^{\frac{1}{3}})$.*

We then walk through the proof highlighting the main Lemmas and steps of the proof.

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 $\text{rad} = \sqrt{2\log(T)/m}$ 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

$$\mathcal{E}_i = \bigcap_{S \cup \{a\} \in \mathcal{S}_i} \{|\bar{f}(S \cup \{a\}) - f(S \cup \{a\})| < \text{rad}\}. \tag{5}$$

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 that $\mathcal{E}$ happens with high probability. We use the concentration of empirical means, Equation (5), and properties of submodularity to show the following important lemma.

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

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

Proof. Recall that a_i, defined in (4), 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 $a \in \mathcal{A}_i \setminus S^{(i-1)}$, 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 (5) to two specific arms, the empirically best a_i out of $\mathcal{A}_i$ and the statistically best $a_i^\star$ out of $\Omega \setminus S^{i-1}$.

$$\begin{aligned}
f(S^{(i)}) &= f(S^{(i-1)} \cup \{a_i\}) && \text{(by design)}\\
&\geq \bar{f}(S^{(i-1)} \cup \{a_i\}) - \mathrm{rad} && \text{(using (5))}\\
&\geq \bar{f}(S^{(i-1)} \cup \{a_i^\star\}) - \mathrm{rad} && (a_i \text{ has the highest empirical mean})\\
&\geq f(S^{(i-1)} \cup \{a_i^\star\}) - 2\mathrm{rad}. && \text{(using (5))}
\end{aligned}$$

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

$$f(S^{(i-1)} \cup \{a_i^\star\}) - f(S^{(i-1)}) \geq \frac{1-\varepsilon}{k}(f(S^\star) - f(S^{(i-1)}) \tag{6}$$

Combining the above results, we conclude that

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

□

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,*

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

We use the above Corollary 1 to bound the expected cumulative regret of our proposed algorithm. We split the expected $(1 - \frac{1}{e})$-regret (2) conditioned on the clean event $\mathcal{E}$ into two parts, one for the exploration and one for the exploitation,

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{t=1}^{T}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S_t)]\right)\\
&= \underbrace{\sum_{i=1}^{k}\sum_{t=T_{i-1}+1}^{T_i}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S_t)]\right)}_{\text{Exploration}} + \underbrace{\sum_{t=T_k+1}^{T}\left((1-\frac{1}{e})f(S^\star) - \mathbb{E}[f(S^{(k)})]\right)}_{\text{Exploitation}}.
\end{aligned} \tag{7}$$

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 (7)) as

$$\begin{aligned}\sum_{i=1}^{k}\sum_{t=T_{i-1}+1}^{T_i}\left((1-\frac{1}{e})f(S^\star)-\mathbb{E}[f(S_t)]\right) &\leq \sum_{i=1}^{k} ms_i\left((1-\frac{1}{e})f(S^\star)-\mathbb{E}[f(S^{(i-1)})]\right)\\ &\leq ms_1\sum_{i=1}^{k}\left((1-\frac{1}{e})f(S^\star)-\mathbb{E}[f(S^{(i-1)})]\right)\\ &\leq ms_1k. \end{aligned} \tag{8}$$

The last inequality follows from the rewards bounded between $[0,1]$.

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 (7)) by applying Corollary 1 as follows

$$\begin{aligned}\sum_{t=T_k+1}^{T}\left((1-\frac{1}{e})f(S^\star)-\mathbb{E}[f(S^{(k)})]\right) &\leq \sum_{t=T_k+1}^{T}(\varepsilon+2k\mathrm{rad})\\ &\leq T\varepsilon+2kT\mathrm{rad}. \end{aligned} \tag{9}$$

Combining the upper bounds (8) and (9), we get

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq ms_1k+\varepsilon T+2kT\mathrm{rad}. \tag{10}$$

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

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq mn\min\{k,\log(\frac{1}{\varepsilon})\}+\varepsilon T+2kT\sqrt{\frac{\log(T)}{m}}. \tag{11}$$

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

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] \leq mn\log(\frac{1}{\varepsilon})+\varepsilon T+2kT\sqrt{\frac{\log(T)}{m}} \tag{12}$$

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

$$0=-mn\frac{1}{\varepsilon}+T+0 \qquad\Rightarrow\qquad \varepsilon=\frac{mn}{T} \tag{13}$$

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

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

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq mn\log(\frac{T}{mn})+\frac{mn}{T}T+2kT\sqrt{\frac{\log(T)}{m}}\\ &\leq mn\log(T)+mn+2kT\sqrt{\frac{\log(T)}{m}}\\ &\leq 2mn\log(T)+2kT\sqrt{\frac{\log(T)}{m}} \end{aligned} \tag{14}$$

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

$$m^\star=\left(\frac{kT\sqrt{\log(T)}}{2n\log(T)}\right)^{2/3}=\left(\frac{kT}{2n\sqrt{\log(T)}}\right)^{2/3} \tag{15}$$

For $T\geq nk\sqrt{\log(T)/2}$, 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,

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq \lceil m^\star\rceil n\log T+2kT(\log(T))^{1/2}\lceil m^\star\rceil^{-1/2}\\ &\leq 2m^\star n\log T+2kT(\log(T))^{1/2}(m^\star)^{-1/2}\\ &= 2^{\frac{1}{3}}k^{\frac{2}{3}}T^{2/3}n^{-2/3}(\log T)^{-1/3}n\log T+2^{\frac{4}{3}}k^{\frac{2}{3}}T(\log(T))^{1/2}T^{-1/3}n^{1/3}(\log T)^{1/6}\\ &= 2^{\frac{1}{3}}k^{\frac{2}{3}}T^{2/3}n^{1/3}(\log T)^{2/3}+2^{\frac{4}{3}}k^{\frac{2}{3}}T^{2/3}n^{1/3}(\log T)^{2/3}\\ &\leq \mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{2}{3}}). \end{aligned} \tag{16}$$

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

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

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 $\mathbb{E}(R(T) \mid \bar{\mathcal{E}}) \leq T$. Moreover, by using the Hoeffding inequality [Hoeffding, 1994], for $T \geq nk\sqrt{\log(T)/2}$, we have $\mathbb{P}(\bar{\mathcal{E}}) \leq \frac{2}{T^3}$, see Lemma 3 in Appendix A. Therefore, we obtain $\mathbb{E}(R(T)) \leq \mathcal{O}(n^{\frac{1}{3}}k^{\frac{2}{3}}T^{\frac{2}{3}}\log(T)^{\frac{1}{3}})$.

Remark 1. *The framework of Nie et al. [2023] that adapts offline algorithms 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 $(1-\frac{1}{e}-\varepsilon)$-approximation guarantee. Then, their framework only provide $(1-\frac{1}{e}-\varepsilon)$-regret guarantee, i.e., weaker guarantee than what we provide.*

Remark 2. *For unknown time horizon T, the geometric doubling trick can extend our result to an anytime algorithm. To initialise 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 only constant factor changes.*

Remark 3. *For the scenario we study in this paper, with stochastic 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 $(1-\frac{1}{e})$-regret is possible (ignoring n and k dependence), and only $\tilde{\mathcal{O}}(T^{2/3})$ bounds have been shown in the existing literature; see Table 1.*

5 Experiments on Online Social Influence Maximization

5.1 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 example, one possible way to market a newly developed product is by selecting a set of highly influential people and hoping they recommend it to their communities. Influence maximization can be modelled as a monotone submodular maximization problem. Selecting more people would not negatively impact the other nodes but would have diminishing returns. Most works consider that there is a fixed constraint on cardinality. Several works considered the problem 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] which may be available for some public accounts but is not available for most social networks. More recently, Nie et al. [2022] proposed ETCG for influence maximization under full-bandit feedback. As shown empirically, their proposed algorithm outperforms other full-bandit algorithms. Therefore, in this work, we compare our method for different ε values, denoted as SGB(ε), including the optimized $\varepsilon^\star = (\frac{nk^2}{4T\log(T)})^{1/3}$, which is denoted as SGB($\varepsilon^\star$), against ETCG Nie et al. [2022, 2023].

5.2 Experiment Details

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

We experimented with an influence maximization problem 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 diffusion process is simulated using the independent cascade model [Kempe et al., 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 neighbours. We used 0.1 uniform infection probabilities for each edge. For every time horizon $T \in \{2\times10^4, 3\times10^4, 4\times10^4, 5\times10^4\}$, we tested each method ten times.

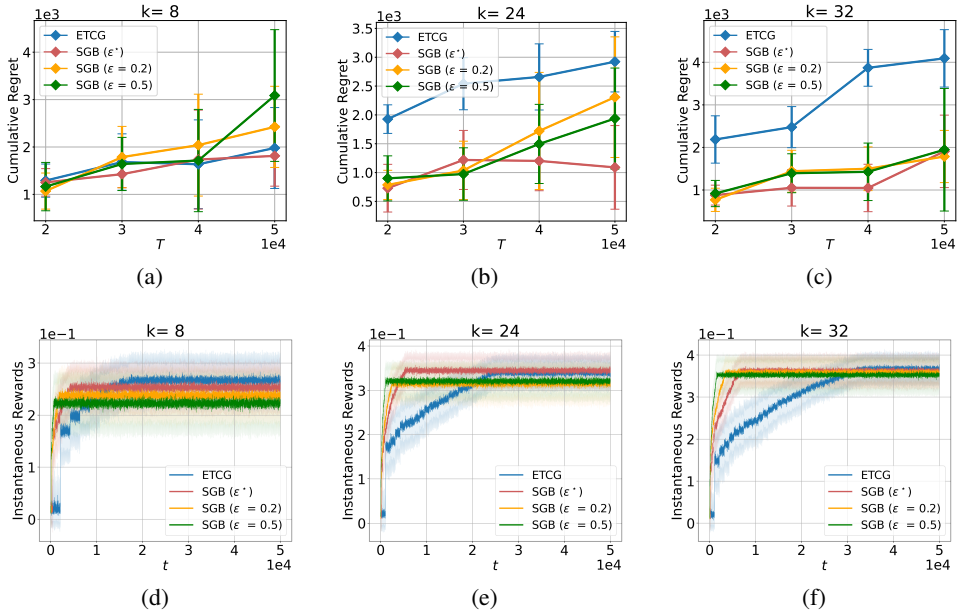

Figure 1: Comparison of SGB, for different ε values, including $\varepsilon^\star = (\frac{nk^2}{4T\log(T)})^{1/3}$, and ETCG. (a), (b), and (c) are the cumulative regret results as a function of horizon T. (d), (e), and (f) show the moving average plots of the immediate rewards as a function of t, with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

5.3 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 $\varepsilon^\star = (\frac{nk^2}{4T\log(T)})^{1/3}$, 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 $\varepsilon^\star$ 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 $\varepsilon^\star$ as shown in Figures (1b) and (1c), respectively, SGB with all the ε values outperforms ETCG with lower average cumulative regrets. Furthermore, Figures (1d), (1e), and (1f) illustrate immediate rewards over a horizon $T = 5 \times 10^4$ for cardinality constraints k of 8, 24, and 32, and $\varepsilon^\star$ 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 converges to a slightly better solution, SGB with all ε converges much faster. For $k = 24$, as shown in Fig. (1e), SGB using $\varepsilon^\star$ converges much faster than ETCG and achieves a better solution. Using other ε values converges 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 converges 30 times faster than ETCG to a solution within a 0.01 neighbourhood of 0.37. Furthermore, using $\varepsilon^\star$ yields the best solution compared to using the other values. Therefore, as predicted by the theory, SGB using $\varepsilon^\star$ has lower cumulative regret than ETCG, and it is more likely to achieve better results than using other values. Additionally, as observed in the experiments and predicted by the theory, our method becomes more effective for larger values of k.

6 Conclusion

We propose the SGB algorithm, which enhances its performance by searching only from a random subset of arms with optimized size. Therefore, SGB has a reduced exploration search space compared to other combinatorial algorithms, making it more efficient for practical problems with many arms n and large cardinality constraint k. Furthermore, we prove that SGB achieves a $(1 - \frac{1}{e})$-regret bound of $\mathcal{O}(n^{\frac{1}{3}} k^{\frac{2}{3}} T^{\frac{2}{3}} \log(T)^{\frac{1}{3}})$, surpassing previous state-of-the-art results.

References

Mridul Agarwal, Vaneet Aggarwal, Christopher J Quinn, and Abhishek K Umrawal. Stochastic top-k subset bandits with linear space and non-linear feedback. In *Algorithmic Learning Theory*, pages 306–339. PMLR, 2021a.

Mridul Agarwal, Vaneet Aggarwal, Abhishek Kumar Umrawal, and Chris Quinn. Dart: Adaptive accept reject algorithm for non-linear combinatorial bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6557–6565, May 2021b.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeff Bilmes. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*, 2022.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *ArXiv*, abs/1803.06971, 2018.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008: 10008, 2008.

Niv Buchbinder, Moran Feldman, Joseph Seffi, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5): 1384–1402, 2015.

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

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

Uriel Feige, Vahab S Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

Fares Fourati, Vaneet Aggarwal, Christopher Quinn, and Mohamed-Slim Alouini. Randomized greedy learning for non-monotone stochastic submodular maximization under full-bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 7455–7471. PMLR, 2023a.

Fares Fourati, Salma Kharrat, Vaneet Aggarwal, Mohamed-Slim Alouini, and Marco Canini. FilFL: Accelerating federated learning via client filtering. *arXiv preprint arXiv:2302.06599*, 2023b.

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

Daniel Golovin, Andreas Krause, and Matthew Streeter. Online submodular maximization under a matroid constraint with application to learning assignments. *arXiv preprint arXiv:1407.1082*, 2014.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 100–108, Bejing, China, 22–24 Jun 2014. PMLR.

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

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

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*, pages 137–146, 2003.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Siyu Lei, Silviu Maniu, Luyi Mo, Reynold Cheng, and Pierre Senellart. Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654, 2015.

Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Shuai Li, Fang Kong, Kejie Tang, Qizhi Li, and Wei Chen. Online influence maximization under linear threshold model. *arXiv preprint arXiv:2011.06378*, 2020.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Rad Niazadeh, Negin Golrezaei, Joshua Wang, Fransisca Susan, and Ashwinkumar Badanidiyuru. Online learning via offline greedy: Applications in market design and optimization. *EC 2021, Management Science Journal*, 2020.

Rad Niazadeh, Negin Golrezaei, Joshua R Wang, Fransisca Susan, and Ashwinkumar Badanidiyuru. Online learning via offline greedy algorithms: Applications in market design and optimization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 737–738, 2021.

Guanyu Nie, Mridul Agarwal, Abhishek Kumar Umrawal, Vaneet Aggarwal, and Christopher John Quinn. An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Guanyu Nie, Yididiya Y Nadew, Yanhui Zhu, Vaneet Aggarwal, and Christopher John Quinn. A framework for adapting offline algorithms to solve combinatorial multi-armed bandit problems with bandit feedback. *arXiv preprint arXiv:2301.13326*, 2023.

Pierre Perrault, Jennifer Healey, Zheng Wen, and Michal Valko. Budgeted online influence maximization. In *International Conference on Machine Learning*, pages 7620–7631. PMLR, 2020.

Lijing Qin and Xiaoyan Zhu. Promoting diversity in recommendation by entropy regularizer. In *IJCAI*, 2013.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Tim Roughgarden and Joshua R. Wang. An optimal learning algorithm for online unconstrained submodular maximization. In *Proceedings of the 31st Conference On Learning Theory*, pages 1307–1325, 2018.

Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, page 1577–1584, Red Hook, NY, USA, 2008. Curran Associates Inc.

Sho Takemori, Masahiro Sato, Takashi Sonoda, Janmajay Singh, and Tomoko Ohkuma. Submodular bandit problem under multiple constraints. In *Conference on Uncertainty in Artificial Intelligence*, pages 191–200. PMLR, 2020.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks VS Lakshmanan, and Mark Schmidt. Model-independent online learning for influence maximization. In *International Conference on Machine Learning*, pages 3530–3539. PMLR, 2017.

Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3026–3036, 2017.

Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. *Advances in Neural Information Processing Systems*, 24, 2011.

A Lemmas and Proofs

Proof of Corollary 1.

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

$$f(S^{(k)}) \geq \left[\frac{1-\varepsilon}{k} f(S^\star) - 2\mathrm{rad}\right] \sum_{\ell=0}^{k-1} (1 - \frac{1-\varepsilon}{k})^\ell. \tag{17}$$

Simplifying the geometric summation,

$$\begin{aligned}\sum_{\ell=0}^{k-1} (1 - \frac{1}{k})^\ell &= \frac{1 - (1 - \frac{1-\varepsilon}{k})^k}{1 - (1 - \frac{1-\varepsilon}{k})} \\ &= k\left(1 - \left(1 - \frac{1-\varepsilon}{k}\right)^k\right).\end{aligned}$$

Continuing with (17),

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

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

$$\begin{aligned}f(S^{(k)}) &\geq (1 - \frac{1}{e} - \varepsilon) f(S^\star) - 2k\mathrm{rad} && (18) \\ &= (1 - \frac{1}{e}) f(S^\star) - \varepsilon f(S^\star) - 2k\mathrm{rad} && \text{(rearranging)} \\ &\geq (1 - \frac{1}{e}) f(S^\star) - \varepsilon - 2k\mathrm{rad} && (f(S^\star) \leq 1) \\ &= (1 - \frac{1}{e}) f(S^\star) - (\varepsilon + 2k\mathrm{rad})\end{aligned}$$

□

Lemma 2 (Hoeffding's inequality)**.** *Let $X_1, \cdots, X_n$ be independent random variables bounded in the interval $[0,1]$, and let $\bar{X}$ denote their empirical mean. Then we have for any $\gamma > 0$,*

$$\mathbb{P}\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \gamma\right) \leq 2\exp\left(-2n\gamma^2\right). \tag{19}$$

Lemma 3. *The probability of the clean event $\mathcal{E}$, for $T \geq kn\sqrt{\log(T)}$, satisfies:*

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2}{T^3}.$$

Proof. We begin by breaking up the probability of the clean event $\mathcal{E}$ into conditional probabilities for the events $\{\mathcal{E}_i\}_{i=1}^k$ for each phase,

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_k) \\ &= \prod_{i=1}^{k} \mathbb{P}(\mathcal{E}_i | \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}).\end{aligned} \tag{20}$$

Recall that $\mathcal{E}_i$ is the event where the empirical means of all actions played in phase i were concentrated around their statistical means. Which actions are available in phase i, namely $\{S^{(i-1)} \cup \{a\}\}_{a \in \mathcal{R} \setminus S^{(i-1)}}$, depends on the action $S^{(i-1)}$ from the previous phase that had the highest empirical mean, which in turn is related to $\mathcal{E}_{i-1}$. Although we cannot directly evaluate (20), by conditioning on $S^{(i-1)}$ we will be able to obtain a bound on (20).

$$\begin{aligned}\mathbb{P}(\mathcal{E}_i | \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) &= \sum_{S \in \left\{S' \,\middle|\, S' \subseteq \Omega,\ |S'| = i-1\right\}} \mathbb{P}(S^{(i-1)} = S, \mathcal{E}_i | \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\ &\qquad\qquad \text{(law of total probability)} \\ &= \sum_{S \in \left\{S' \,\middle|\, S' \subseteq \Omega,\ |S'| = i-1\right\}} \mathbb{P}(S^{(i-1)} = S | \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \mathbb{P}(\mathcal{E}_i | S^{(i-1)} = S, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\ &= \sum_{S \in \left\{S' \,\middle|\, S' \subseteq \Omega,\ |S'| = i-1\right\}} \mathbb{P}(S^{(i-1)} = S | \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \mathbb{P}(\mathcal{E}_i | S^{(i-1)} = S),\end{aligned} \tag{21}$$

where (21) follows from rewards in phase i being conditionally independent of rewards from other phases, given the corresponding actions played during phase i.

We now focus on bounding $\mathbb{P}(\mathcal{E}_i | S^{(i-1)} = S)$. By conditioning on the set chosen in the previous phase, $S^{(i-1)} = S$, we know all the actions that will be played in the current phase i, $\{S^{(i-1)} \cup \{a\}\}_{a \in \mathcal{R} \setminus S^{(i-1)}}$. The rewards of all the actions are bounded in $[0,1]$ and are conditionally independent (given the corresponding action).

Apply Lemma 2 to the empirical mean $\bar{f}(S^{(i-1)} \cup \{a\})$ of m rewards for action $S^{(i-1)} \cup \{a\}$ and choosing $\epsilon = \text{rad} = \sqrt{2\log(T)/m}$ gives

$$\begin{aligned}\mathbb{P}\left[\left|\bar{f}(S^{(i-1)} \cup \{a\}) - f(S^{(i-1)} \cup \{a\})\right| \geq \text{rad}\right] &\leq 2\exp\left(-2m\text{rad}^2\right) \\ &= 2\exp\left(-2m(2\log(T)/m)\right) \\ &= 2\exp\left(-4\log(T)\right) \\ &= \frac{2}{T^4}.\end{aligned}$$

Thus, for any individual action $S^{(i-1)} \cup \{a\} \in \mathcal{S}_i$, we can bound the probability that its sample mean $\bar{f}(S^{(i-1)} \cup \{a\})$ is within a specified confidence radius (complementary of the event above) as

$$\begin{aligned}\mathbb{P}\left[\left|\bar{f}(S^{(i-1)} \cup \{a\}) - f(S^{(i-1)} \cup \{a\})\right| < \text{rad}\right] &= 1 - \mathbb{P}\left[\left|\bar{f}(S^{(i-1)} \cup \{a\}) - f(S^{(i-1)} \cup \{a\})\right| \geq \text{rad}\right] \\ &\geq 1 - \frac{2}{T^4}.\end{aligned} \tag{22}$$

We can then use (22) to bound $\mathbb{P}(\mathcal{E}_i|S^{(i-1)} = S)$ for any set $S \subset \Omega$ of $i-1$ arms.

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_i|S^{(i-1)} = S) &= \mathbb{P}\left[\bigcap_{a\in\mathcal{R}\setminus S^{(i-1)}} \left\{\left|\bar{f}(S^{(i-1)} \cup \{a\}) - f(S^{(i-1)} \cup \{a\})\right| < \text{rad}\right\} \middle| S^{(i-1)} = S\right] \\
&\qquad\qquad \text{(definition of } \mathcal{E}_i\text{)} \\
&= \prod_{a\in\mathcal{R}\setminus S^{(i-1)}} \mathbb{P}\left[\left\{\left|\bar{f}(S^{(i-1)} \cup \{a\}) - f(S^{(i-1)} \cup \{a\})\right| < \text{rad}\right\} \middle| S^{(i-1)} = S\right] \\
&\qquad\qquad \text{(rewards are independent conditioned on actions)} \\
&\geq \left(1 - \frac{2}{T^4}\right)^{|\mathcal{R}\setminus S^{(i-1)}|} \qquad \text{(using (22))} \\
&= \left(1 - \frac{2}{T^4}\right)^{n \min\{1, \frac{\log(\frac{1}{\varepsilon})}{k}\} - i + 1} \\
&\geq \left(1 - \frac{2}{T^4}\right)^{n}. \qquad (23)
\end{aligned}$$

Using (21) and (23), we are now ready to lower bound the probability of a clean event.

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_k) \\
&= \prod_{i=1}^{k} \mathbb{P}(\mathcal{E}_i|\mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\
&= \prod_{i=1}^{k} \sum_{S\in\left\{S' \,\middle|\, S'\subseteq\Omega,\ |S'|=i-1\right\}} \mathbb{P}(S^{(i-1)} = S|\mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \mathbb{P}(\mathcal{E}_i|S^{(i-1)} = S) \\
&\qquad\qquad \text{(using (21))} \\
&\geq \prod_{i=1}^{k} \sum_{S\in\left\{S' \,\middle|\, S'\subseteq\Omega,\ |S'|=i-1\right\}} \mathbb{P}(S^{(i-1)} = S|\mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \left(1 - \frac{2}{T^4}\right)^{n} \qquad \text{(using (23))} \\
&= \prod_{i=1}^{k} \left(1 - \frac{2}{T^4}\right)^{n} \sum_{S\in\left\{S' \,\middle|\, S'\subseteq\Omega,\ |S'|=i-1\right\}} \mathbb{P}(S^{(i-1)} = S|\mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\
&= \prod_{i=1}^{k} \left(1 - \frac{2}{T^4}\right)^{n} \\
&= \left(1 - \frac{2}{T^4}\right)^{nk} \\
&\geq 1 - \frac{2nk}{T^4}. \qquad \text{(Bernoulli's inequality)}
\end{aligned}$$

For $T \geq nk\sqrt{\log(T)}$, we have $T \geq nk$, therefore

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2}{T^3}. \qquad (24)$$

This concludes the proof for Lemma 3. □