Incentivized Bandit Learning with Self-Reinforcing User Preferences

Anonymous Author
Anonymous Institution

Abstract

In this paper, we propose to study a new multi-armed bandit (MAB) online learning model that combines two real-world phenomena in many recommender systems in practice: (i) the service provider cannot pull the arms by itself and thus has to offer rewards to users to incentivize arm-pulling indirectly; and (ii) if users with specific arm preferences are well rewarded, they induce a "self-reinforcing" effect in the sense that they will attract more users of similar arm preferences. The goal of the service provider is to maximize the total rewards over a time horizon T with a low total payment. Our contributions in this paper are two-fold: (i) We propose a new MAB model that considers both users' self-reinforcing preference behaviors and incentives; and (ii) We leverage the properties of a multi-color Pólya urn model with nonlinear feedback to propose two policies termed "At-Least-n Greedy" and "UCB-List." We prove that both policies achieve $O(\log T)$ expected regrets with $O(\log T)$ expected payments over a time horizon T. We conduct extensive experiments to demonstrate the performances of these two policies and compare their robustness under various settings.

1 INTRODUCTION

In many online learning systems, there exists a self-reinforcing phenomenon, where the current user's behavior is influenced by the user behaviors in the past (Barabási and Albert, 1999; Chakrabarti et al., 2005; Ratkiewicz et al., 2010), or an item is getting increasingly more popular as it accumulates more positive feedbacks. For example, on a movie rental website, current customers tend to have more interest in Movie A that has 500 positive reviews, compared with Movie B that only has 10 positive reviews. On the other hand, the website owner, who aims to maximize the total profit in the long run, wants to identify the most profitable movies. The online learning problem under self-reinforcing preferences can be modeled by the multi-armed bandit (MAB) framework (Berry and Fristedt, 1985; Bubeck and Cesa-Bianchi, 2012). In the classic stochastic MAB setting, an agent observes and chooses to pull an arm (action). The chosen arm, as feedback, generates a random reward from an unknown distribution and is assumed to be independent and identically distributed (i.i.d.) over time. The agent, aiming to maximize total rewards over a given time horizon T, tries to design a strategy to perform arm selection in the presence of unknown arm reward distributions.

In the literature, existing works (Fiez et al., 2018; Shah et al., 2018) that incorporate the self-reinforcing preferences into the MAB model remain limited. Shah et al. (2018) showed that the self-reinforcing preferences might render the classic UCB (upper confidence bound) algorithm (Auer et al., 2002) sub-optimal, and proposed new optimal arm selection algorithms. However, in many platforms that utilize the MAB framework for online sequential decision making (e.g., recommender systems, healthcare, finance, and dynamic pricing, see Bouneffouf and Rish (2019)), the service providers cannot select arms directly. Rather, arms are pulled by users who are exhibiting self-reinforcing preferences. The service provider thus needs to incentivize users to select certain arms to maximize the total rewards, while avoiding incurring high incentive costs. Hence, the bandit models in Fiez et al. (2018); Shah et al. (2018) are no longer applicable even though the self-reinforcing preferences behavior is considered. On the other hand, there exist several works (Frazier et al., 2014; Mansour et al., 2015, 2016; Wang and Huang, 2018) that studied incentivized bandit under various settings and proposed efficient algorithms (more details in Section 2), but none of these works models users with self-reinforcing preferences. To our knowl-

Preliminary work. Under review by AISTATS 2021. Do not distribute.

edge, our work is the first to jointly consider both incentives and self-reinforcing preferences in the MAB framework – two key features of many online learning systems in practice. As will be seen later, the combined effect of incentives and self-reinforcing preferences creates significant challenges in analyzing this new MAB model.

Specifically, in this paper, we propose a new MAB model to address the self-reinforcing preferences and incentivized arm selections. In this model, two fundamental trade-offs naturally emerge: On the one hand, sufficient exploration is required to find an optimal arm, which may result in pulling sub-optimal arms, while adequate exploitation is needed to stick with the arm that did well in the past, which may or may not be the best choice in the long run. On the other hand, the service provider needs to provide enough incentives to mitigate unfavorable self-reinforcing preferences, while in the meantime avoiding unnecessarily high compensations for users. As in most online learning problems, we use *regret* as a benchmark to evaluate the performance of our learning policy, which is defined as the performance gap between the proposed policy and an optimal policy in hindsight. The major challenges thus lie in three aspects: (a) During incentivized pulling, how could the service provider maintain a balance between exploration and exploitation to minimize regret? (b) How long should the service provider incentivize arm selections until the right self-reinforcing preferences are established toward an optimal arm, so that no further incentive is needed? (c) Is the self-reinforcing preferences strong and stable enough to sustain the sampling of an optimal arm over time without additional incentives? If yes, under what conditions could this happen?

In this work, we address the above challenges and questions by proposing two optimal policies for the incentivized MAB framework with self-reinforcing preferences. Our main contributions are as follows:

- We propose an incentivized bandit learning model with nonlinear feedback. Compared to the basic stochastic MAB model, this proposed model considers non-deterministic pulling from users who exhibit self-reinforcing preferences, and the service provider implements a bandit policy with incentives. To our knowledge, this is the first work that integrates both self-reinforcing preference and incentive in MAB.
- We show that no incentivized bandit policy achieves a sub-linear regret with a sub-linear total payment if the feedback function that models the self-reinforcing preferences does not satisfy certain conditions. The proof is inspired by a multi-color Pólya urn model, and we also show how to guide the self-reinforcing preferences toward a desired direction.
- We propose two bandit policies, namely At-Least-n Greedy and UCB-List, both of which are optimal in regret. Specifically, for the two policies, we analyze the upper bounds of the expected regret and the expected total payment over a fixed time horizon T. We show that both policies achieve $O(\log T)$ expected regrets, which meet the lower bound in Lai and Robbins (1985). Meanwhile, the expected total incentives for both policies are upper bounded by $O(\log T)$.
- We conduct extensive simulations to demonstrate the performance of both policies. Our results show that both policies are not only effective in performance but also robust under various settings.

2 RELATED WORK

The self-reinforcing phenomenon has received increasing interest in several different fields recently. In the random network literature, previous works have studied the network evolution with "preferential attachment" (Barabási and Albert, 1999; Chakrabarti et al., 2005; Ratkiewicz et al., 2010). Also, a similar social behavior, referred to as *herding*, is studied in the Bayesian learning model literature (Bikhchandani et al., 1992; Smith and Sørensen, 2000; Acemoglu et al., 2011). For example, Acemoglu et al. (2011) first studied the conditions under which there exists a convergence in probability to the desired action as the size of a social network increases. More recently, Shah et al. (2018) incorporated positive externalities in user arrivals and proposed bandit algorithms to maximize the total reward. Then, Fiez et al. (2018) provided a more general model, where the service provider has limited information. Unlike previous works, the service providers in Shah et al. (2018); Fiez et al. (2018) have full control in determining which arm for users to pull. In contrast, the service provider in our model can only incentivize users to indirectly induce the preferences toward a desired direction, and which arm to be pulled is entirely dependent on the current user's random preference.

On the other hand, incentivized MAB has attracted growing attention in recent years (Kremer et al., 2014; Frazier et al., 2014; Mansour et al., 2015, 2016; Wang and Huang, 2018). To our knowledge, Frazier et al. (2014) first adopted incentive schemes into a Bayesian MAB setting. In their model, the service provider seeks to maximize time-discounted total rewards by incentivizing arm selections. Kremer et al. (2014) shares a similar motivation as Frazier et al. (2014). But in the model of Kremer et al. (2014), the service provider does not offer payments to the users. Instead, he decides the information to be revealed to users as incentives. Subsequently, Mansour et al. (2015) studied the case where the rewards are not discounted over time. More recently, Wang and Huang (2018) considered the non-Bayesian setting with non-discounted rewards. These models differ from our model in both the incentive schemes and user behaviors. Another line of research similar to incentivized bandit is bandit with budgets (Guha and Munagala, 2007; Goel et al., 2009; Combes et al., 2015; Xia et al., 2015), where the service provider takes actions with budget constraints. Guha and Munagala (2007) developed approximation algorithms for a large class of budgeted learning problems. Then, Goel et al. (2009) proposed index-based algorithms for this problem. The key difference from our work is that in these models, the budget constraints are pre-determined, and the service providers cannot take any further actions as soon as the budget constraints are violated. In contrast, the total payment in our model is evaluated only after the time horizon is finished, which implies that bounding the total payment is also part of our goals.

Although not cast in the MAB framework, the works on *urn models* (Khanin and Khanin, 2001; Drinea et al., 2002; Oliveira, 2009; Zhu, 2009) also share some relevant feedback settings to our model. Drinea et al. (2002) first proposed a class of processes called *balls and bins models with feedback*, which is a preferential attachment model for large networks. They then proved the convergence results of the model with various feedback functions. Later, Khanin and Khanin (2001) improved the convergence result by showing monopoly (to be defined later) happens with probability one under a class of feedback functions included in Drinea et al. (2002). Our proposed model is inspired by the ideas of feedback from Oliveira (2009), in which the author discussed a natural evolution of the balls and bins process with nonlinear feedback. However, our model is focused on MAB regret minimization, which is completely different from the goals considered in these works.

3 MODELING AND NOTATIONS

In this paper, we denote the set of arms offered by the service provider as $A = \{1, \ldots, m\}$. Each arm a follows a Bernoulli reward distribution D_a with an unknown mean $\mu_a > 0$. The process runs for T rounds. At each time step $t \in \{1, \ldots, T\}$, a user arrives and chooses an arm $I(t)$ to pull. After pulling arm $I(t)$, the service provider receives a random reward $X(t) \sim D_{I(t)}$. For $t \geq 1$, we denote the number of times that an arm a is pulled up to time t as $T_a(t) := \sum_{i=1}^{t} \mathbb{1}_{\{I(i)=a\}}$, and denote the total reward generated by arm a up to time t as $S_a(t) := \sum_{i=1}^{t} X(i) \cdot \mathbb{1}_{\{I(i)=a\}}$. For each arm $a \in A$, we let $T_a(0) = 0$ and $S_a(0) = 0$. We assume that there exists a unique best arm $a^* \in A$ such that $a^* = \arg\max_a \mu_a$.

1) Preference and bias modeling: In our model, the user behavior is stochastic. Specifically, in each time step t, the user has a positive probability $\lambda_a(t)$ to pull each arm $a \in A$, with $\sum_{a \in A} \lambda_a(t) = 1, \forall t$. In other words, the probability $\lambda_a(t)$ can be viewed as the preference rate in time step t. In this paper, we adopt the widely used multinomial logit model in the literature to model $\lambda_a(t)$ as follows:

$$\lambda_a(t) = \frac{F\big(S_a(t-1) + \theta_a\big)}{\sum_{i \in A} F\big(S_i(t-1) + \theta_i\big)}, \qquad (1)$$

where the function $F : \mathbb{R} \to (0, +\infty)$ is an increasing feedback function of the accumulative reward of an arm a, and $\theta_a > 0$ denotes the initial preference bias of arm a. Several important remarks for the preference model in (1) are in order: (i) The multinomial logit model is based on the behavioral theory of utility and has been widely applied in the marketing literature to model the brand choice behavior (Guadagni and Little, 2008; Gupta, 1988). The multinomial logit model is also used in the social network literature to model preferential attachment (Barabási and Albert, 1999), where the probability that a link connects a new node j with another existing node i is linearly proportional to the degree of i. Besides, the preferential attachment phenomenon also characterizes user behavior influenced by history. Therefore, we assume in (1) that the user can access the history and will be influenced accordingly. (ii) For the feedback function $F(\cdot)$ in (1), a simple example is $F(x) = x^\alpha$ for some $\alpha > 1$, i.e., users are more influenced by the accumulative reward in the past with a larger α-value.

2) Incentive mechanism modeling: In the classic MAB model, the service provider can directly control which arm to pull (Shah et al., 2018). In our model, however, the user randomly selects an arm depending on the current preference rate, which is in turn influenced by the history and the incentive. The service provider aims at maximizing total reward in the long run, but cannot directly control which arm to pull. The service provider can only offer a certain amount of incentive on the arm that he prefers, so as to increase the users' preferences of pulling this particular arm. In this paper, we model the influence of the incentives following the so-called "coupon effects on brand choice behaviors" in the economics literature (Papatla and Krishnamurthi, 1996; Bawa and Shoemaker, 1987). Simply speaking, the relationship between coupons and choices is nonlinear, and the redemption rate increases with respect to the coupon face value but exhibits a diminishing return effect (Bawa and Shoemaker, 1987). This type of incentive effects can be modeled as follows: in each time step t, if arm $a \in A$ is preferred by the service provider, a payment b is offered to the user to increase the user's preference on pulling arm a. The preference rates under payment b are updated as follows:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{G'(b) + F\big(S_i(t-1) + \theta_i\big)}{G'(b) + \sum_{j \in A} F\big(S_j(t-1) + \theta_j\big)}, & i = a, \\ \dfrac{F\big(S_i(t-1) + \theta_i\big)}{G'(b) + \sum_{j \in A} F\big(S_j(t-1) + \theta_j\big)}, & i \neq a, \end{cases}$$

where $G'(\cdot)$ is an increasing function independent of $S_i(t-1)$, F and θ_i. Clearly, the above preference update model still follows the multinomial logit model, where the value $G'(b)$ can be interpreted as the perceived reward to the users for pulling arm a under payment b. We assume that the service provider has the knowledge of $G'(\cdot)$, which can be learned from historical data. Also, we can see that from the above definition that, as the payment b increases asymptotically (i.e., $b \uparrow \infty$), we have $\hat{\lambda}_a(t) \uparrow 1$ and $\hat{\lambda}_i(t) \downarrow 0$, $\forall i \neq a$, i.e., arm a is preferred with probability one.

For simplicity and convenience in our subsequent theoretical analysis, in the rest of the paper, we rewrite $\hat{\lambda}_i(t)$ in the following equivalent form: we divide both the denominator and numerator by $\sum_{i \in A} F\big(S_i(t-1) + \theta_i\big)$ and let $G(b) \triangleq G'(b) / \sum_{i \in A} F\big(S_i(t-1) + \theta_i\big)$. Then, it can be readily verified that the updated preference rate can be equivalently rewritten as:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{\lambda_i(t) + G(b)}{1 + G(b)}, & i = a, \\ \dfrac{\lambda_i(t)}{1 + G(b)}, & i \neq a. \end{cases}$$

It is easy to see that $G(\cdot)$ remains an increasing function of b. Also for notational convenience, we define the accumulative payment offered up to time step t as $B_t \coloneqq \sum_{i=1}^{t} b_t$, where $b_t \in \{0, b\}$, $\forall t$.

3) Regret: Let $\Gamma_T = \sum_{t=1}^{T} X(t)$ denote the accumulative rewards up to time T. In this paper, we aim to maximize $\mathbb{E}[\Gamma_T]$ by designing an incentivized policy π with low accumulative payment, hopefully, both being logarithmic with respect to the time horizon T. A policy π is an algorithm that produces a sequence of arms that are recommended at time step $t = 1, \ldots, T$. Similar to classic MAB problems, we want to measure our performance against an oracle policy, where in hindsight the service provider knows the best arm a^* with the largest mean and can always offer an *infinite* amount of payments to users, so that the updated preference rate of arm a^* is always infinitely close to one. We denote the expected accumulative reward generated under the oracle policy up to time T as $\mathbb{E}[\Gamma_T^*] = \mu_{a^*} T$.[1] The expected (pseudo) regret is defined as follow: $\mathbb{E}[R_T] = \mu_{a^*} T - \mathbb{E}[\Gamma_T]$. Our objective is to minimize the expected regret $\mathbb{E}[R_T]$, with the expected accumulative payment $\mathbb{E}[B_T]$ being logarithmic with respect to the time horizon T.

4 POLICIES AND UPPER BOUNDS

In this section, we present two policies that achieve $O(\log T)$ expected regret with $O(\log T)$ accumulative payment with respect to the time horizon T. The main ideas of these two policies are similar. We first perform exploration among all arms by incentivizing pulling until we know the best-empirical arm is optimal, i.e., $\hat{a}^* = a^*$ with high confidence. Then, we keep incentivizing the pulling of the best-empirical arm $\hat{a}^*$ until it dominates. To this end, we formally define the notion of dominance as follows:

Definition 1 (Dominance)**.** *An arm is said to be dominant if it produces at least half of the total reward.*

The key of the policies is to guarantee the dominance of arm $\hat{a}^*$, which induces the correct preference to arm $\hat{a}^*$, while keeping the accumulative payment sub-linear.

Before presenting the policies, we want to show that if the feedback function $F(x)$ satisfies certain conditions, then with probability one, the users' self-reinforcing preferences converge to one arm without incentive after sufficiently many time steps, i.e., an arm $a \in A$ is the only arm to be sampled. We define this event as the *monopoly by arm a*, or *mono*$_a$ for short. We then define an incentivized policy as a policy that incentivizes pulling with bounded payment for each time step for a continuous duration smaller than or equal to the time horizon T. Clearly, monopoly is necessary for the existence of an incentivized policy that induces all users' preferences to a certain arm with sub-linear to-

[1] It is insightful to compare our oracle policy with Shah et al. (2018). The oracle policy in Shah et al. (2018) does not achieve $\mu_{a^*} T$ expected accumulative reward up to time T due to the following key modeling difference: In Shah et al. (2018), it is assumed that the service provider can only feed a *single arm* at a time to the current user. Hence, the oracle policy keeps *only* feeding the best arm to all arriving users. However, in the early time steps, a fraction of the users may not prefer the best arm due to initial biases. Hence, the system has to spend time mitigating these initial biases, resulting in an expected accumulative reward smaller than $\mu_{a^*} T$. In contrast, we assume that the service provider can feed *all arms* to each user (closely models real-world recommender systems), and the oracle policy offers an infinite amount of payment as incentives. Thus, users will always pull the best arm with probability one in each time step, which implies $\mu_{a^*} T$ expected accumulative reward up to time T.

tal payment. Correspondingly, if such a policy exists, then it implies that the system enjoys the property that the users' self-reinforcing preferences converge to a certain arm even after the service provider stops providing incentives at some stage.

Lemma 1. (Monopoly) *There exists an incentivized policy that induces users' preferences to converge in probability to an arm over time with sub-linear payment, if and only if $F(x)$ satisfies $\sum_{i=1}^{+\infty} \frac{1}{F(i)} < +\infty$.*

Proof Sketch. Our main mathematical tool is the improved exponential embedding method. In a nutshell, this method simulates the reward generating sequence by random exponentials. Define a sequence $\{\chi_j\}_{j=1}^{\infty}$ denoting the reward generating order, where each element denotes the arm index. Note that an arm index appears in $\{\chi_j\}$ only when it is pulled and generates a unit reward. We want to construct a sequence $\{\zeta_j\}$ that has the same conditional distribution as $\{\chi_j\}$ given history $\mathcal{F}_{j-1}$. For the arm $i \in A$, consider a collection of independent exponential random variables $\{r_i(n)\}$ such that $\mathbb{E}[r_i(n)] = \frac{1}{\mu_i F(n+\theta_i)}$. We construct an infinite set $B_i = \{\sum_{k=0}^{n} r_i(k)\}_{n=0}^{\infty}$, where each element $\sum_{k=0}^{n} r_i(k)$ models the time needed for arm i to get n accumulative rewards. Then we mix up and sort B_i for all $i \in A$ to form a sequence H. Our objective sequence $\{\zeta_j\}$ is the arm index sequence out of H.

Next, we prove by induction that given the previous reward history $\mathcal{F}_{j-1}$, the constructed sequence $\{\zeta_j\}$ has the same conditional distribution as $\{\chi_j\}$. Then, the proof of Lemma 1 is done if we show that if and only if any feedback function $F(x) > 0$ satisfies $\sum_i \frac{1}{F(i)} < +\infty$, then $\mathbb{P}(\exists a \in A, mono_a) = 1$. To prove this, we define the *attraction time* N as the time step when the monopoly happens. By leveraging the constructed sequence $\{\zeta_j\}$, we establish the necessity by showing that if $\sum_i \frac{1}{F(i)} < +\infty$ then $\mathbb{P}(N < \infty) = 1$, and the sufficiency by showing that if $\sum_i \frac{1}{F(i)} = +\infty$ then $\mathbb{P}(N = \infty) > 0$. □

Remark 1. The exponential embedding technique has been widely applied (Zhu, 2009; Oliveira, 2009; Davis, 1990; Athreya and Karlin, 1968). This method embeds a discrete-time process into a continuous-time process built with exponential random variables. We improve this method and adapt it to our model, which considers more random variables. The most significant feature of the exponential embedding technique is that the random times of different arms generating unit rewards are independent and can be mathematically expressed using exponential distributions, which facilitates our subsequent analysis.

Remark 2. As a special case, if $F(x) = x^{\alpha}$ with $\alpha > 1$ (i.e., superlinear polynomial), then there exists an incentivized policy that induces all preferences to converge over time with sub-linear total payment, since $\sum_{i=1}^{+\infty} \frac{1}{i^{\alpha}} < +\infty$ with $\alpha > 1$. Previous works (Drinea et al., 2002; Khanin and Khanin, 2001) considering the balls and bins model also studied the cases $\alpha < 1$ and $\alpha = 1$. In the case $\alpha < 1$, the asymptotic preference rates of arms are all deterministic, positive, and dependent on the means and biases of arms. In the case $\alpha = 1$, the system is akin to a standard Pólya urn model, and will converge to a state where all arms have random positive preference rates that depend on the means and initial biases of the arms. In the case $\alpha > 1$, the system converges almost surely to a state where only one arm has a positive probability to generate rewards, depending on the means and initial biases of arms. Thus, the system under these three conditions exhibits completely different behaviors.

4.1 At-Least-n Greedy

The At-Least-n Greedy policy consists of three phases: the exploration phase, the exploitation phase, and the self-sustaining phase. The service provider participates in the first two phases. During the exploration phase, the At-Least-n Greedy policy explores all arms until each arm generates sufficient accumulative rewards. Then, the policy exploits the arm with the best empirical mean until it dominates (as defined in Definition 1). We show that once the best-empirical arm dominates, there is an overwhelming probability that after finite time steps in the self-sustaining phase, monopoly happens on the best-empirical arm, which implies sub-linear regret. *One of our key contributions in this paper is the discovery* that the incentive can stop as soon as the dominance happens, which is *earlier than* the monopoly is established and guarantees sub-linear accumulative payment. In both policies, we define the sample mean of arm $a \in A$ in time step t as $\hat{\mu}_a(t) = \frac{S_a(t-1)}{T_a(t-1)}$. The At-Least-$n$ Greedy policy is stated as follows:

Policy 1 (At-Least-n Greedy)**.** *Given the time horizon T, incentive sensitivity function G and a payment b satisfying $G(b) > 1$. Let $n = q \log T$, where $q \geq 1$ is a tunable parameter.*

1) Exploration Phase: *Let $\tau_n = \min(t : S_a(t) \geq n, \forall a) \wedge T$ denote the random time step where any arm has accumulative reward of at least n. For $t \in [1, \tau_n]$, incentivize users with payment b to sample arm $I(t) \in \arg\min_{a \in A} S_a(t)$, with ties broken at random.*

2) Exploitation Phase: *Let $\hat{a}^* \in \arg\max_{a \in A} \hat{\mu}_a(\tau_n + 1)$, with ties broken at random. Let $\tau_s = \tau_n + \delta\tau_n$, where $\delta = \frac{G(b)+1}{G(b)-1}$. For $t \in (\tau_n, \tau_s]$, the service provider offers payment b for users to sample arm $\hat{a}^*$.*

3) Self-Sustaining Phase: *For $t \in (\tau_s, T]$, the service provider stops offering payments and lets users sample arms based on their own preferences.*

Beyond time τ_s, arm $\hat{a}^*$ is expected to dominate. Since with enough time steps after τ_s, monopoly happens with probability one and arm $\hat{a}^*$ has high probability to emerge victorious in monopoly (to be shown in the proof of Theorem 2). If the time horizon T is large enough to cover the attraction time, then arm $\hat{a}^*$ will be sampled repeatedly after monopoly happens, while the accumulative reward generated by sub-optimal empirical arms is *independent* of T after the monopoly (which contribute to the regret). Thus, the policy achieves a sub-linear expected regret stated as follows:

Theorem 2. (At-Least-n Greedy) *Given a sensitivity function $G(\cdot)$ and a fixed payment b satisfying $G(b) > 1$, if the feedback function satisfies $F(x) = \Omega(x^\alpha)$ and $F(x) = O(x^\alpha \ln x)$ for $\alpha > 1$, the expected regret $\mathbb{E}[R_T]$ of the At-Least-n Greedy policy is upper bounded by $\mathbb{E}[R_T] \le \sum_{a\in A} \frac{2\max_{i\in A}\Delta_i + [G(b)-1]\Delta_a}{[G(b)-1]\mu_a} \times q\log T + O(1)$, with the expected accumulative payment $\mathbb{E}[B_T]$ upper bounded by $\mathbb{E}[B_T] \le \sum_{a\in A} \frac{2[G(b)+1]}{[G(b)-1]\mu_a} \times q\log T$.*

Proof Sketch. By the law of total expectation, the expected regret up to time T can be decomposed as $\mathbb{E}[R_T] \le \mathbb{E}[R_T \mid \hat{a}^* = a^*] + T\cdot\mathbb{P}(\hat{a}^* \ne a^*)$. To bound $\mathbb{E}[R_T]$, we want to upper bound both $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$ and $\mathbb{P}(\hat{a}^* \ne a^*)$. First, the probability $\mathbb{P}(\hat{a}^* = a^*)$ is upper bounded by $O(T^{-1})$ by leveraging the Chernoff-Hoeffding bound. To bound $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$, we need to bound $\mathbb{E}[\tau_n]$ and $\mathbb{E}[\tau_s]$. Consider $\mathbb{E}[\tau_n]$, we show that the number of pulling of arm a to get a unit reward is a geometric random variable with parameter larger than $\frac{\mu_a G(b)}{1+G(b)}$. Then, for each arm $a \in A$ to obtain at least n accumulative reward, the expected time needed is upper bounded by $\mathbb{E}[\tau_n] \le n\sum_{a\in A}\frac{1+G(b)}{\mu_a G(b)}$. For $\mathbb{E}[\tau_s]$, it follows from its definition that $\mathbb{E}[\tau_s] = (1+\delta)\mathbb{E}[\tau_n]$. According to the policy, the expected accumulative payment $\mathbb{E}[B_T]$ can be upper bounded by $b\mathbb{E}[\tau_s]$.

The next challenge is to show whether the dominant arm has a large enough probability to "win" in monopoly. By straightforward probability calculations, our choice of τ_s can be shown to be sufficiently large to guarantee that the best-empirical arm dominates in expectation. Next, we show that during the self-sustaining phase, the dominant arm has an overwhelming probability to "win" in monopoly. Construct an event $D(u, n)$ to describe the following phenomenon: the fraction of accumulative reward from a weak arm increases over time, i.e., at time step τ_s, there are $u_0 n_0$ accumulative reward from the weak arm, with n_0 total reward and the fraction $u_0 < \frac{1}{2}$. Then at time step $t' \in (\tau_s, T]$, there are un accumulative reward from the weak arm with the fraction $u > u_0$. The probability of $D(u,n)$ can be bounded as $\mathbb{P}\big(\exists n > n_0,\, D(u,n)\big) \le Ce^{-(u_0 n_0)^\gamma}$ with constant $\gamma > 0$ using the improved exponential embedding method and a Chernoff-like bound developed in Lemma 7. Thus the arms that stay on the weak side for a long time have little chance to win back.

Lastly, with the upper bound of $\mathbb{E}[\tau_n]$ and $\mathbb{E}[\tau_s]$, we obtain $\mathbb{E}[R_T \mid \hat{a}^* = a^*] \le n\sum_{a\in A}\frac{2\max_{i\in A}\Delta_i + [G(b)-1]\Delta_a}{[G(b)-1]\times\mu_a} + \mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$, where $\mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*]$ represents the expected regret during time $(\tau_s, T]$. The expected regret during this phase is caused by pulling any sub-optimal arms, and can be bounded by $O(1)$, which follows from the fact that "the arm that stays on the weak side for a long time has little chance to win back" and some standard expectation calculations. Thus, the term $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$ can be upper bounded by $O(1)$. This completes the proof. □

4.2 UCB-List

In this section, we propose a UCB-List policy to further improve the performance of the At-Least-n Greedy policy. The UCB-List policy is similar to At-Least-n Greedy and also consists of three phases. During the exploration phase, the service provider initially puts all arms in one set, and then makes arm selection in a way similar to the standard UCB policy. Meanwhile, it removes arms that are estimated to be sub-optimal, until only one arm is left in the set, which is viewed as the best-empirical arm. Then, the service provider incentivizes users to sample this arm until it dominates. Compared to At-Least-n Greedy, UCB-List reduces the pulling times of sub-optimal arms during the exploration phase, while still balancing the trade-off between exploration and exploitation. Formally, the UCB-list policy is stated as follows:

Policy 2 (UCB-List)**.** *Given the time horizon T, function $G(\cdot)$ and a payment b satisfying $G(b) > 1$. Define the confidence interval of arm a in time step t as $c_a(t) = p\sqrt{\frac{\ln T}{T_a(t)}}$, where p is a tunable parameter.*

Initialization: *Incentivize users with payment b to sample arm $a \in A$ for which $T_a(t) = 0$, with ties broken at random until $\min_{a\in A} T_a(t) = 1$. Let set $U = A$.*

1) Exploration Phase: *If $|U| > 1$, remove arm a for which $\hat{\mu}_a(t) + c_a(t) \le \min_{i\ne a, i\in U}\big[\hat{\mu}_i(t) - c_i(t)\big]$ from set U if there is any. Offer payment b to incentivize users to sample arm $I(t) \in \arg\max_{a\in U}\big[\hat{\mu}_a(t) + c_a(t)\big]$, with ties broken at random. If $|U| = 1$, let arm $\hat{a}^* = \{a : a \in U\}$, mark current time as τ_1 and proceed to the Exploitation Phase.*

2) Exploitation Phase: *Let $\tau_2 = \tau_1 + \delta\big(\tau_1 - T_{\hat{a}^*}(\tau_1)\big)$, where $\delta = \frac{G(b)+1}{G(b)-1}$. For $t \in [\tau_1, \tau_2]$, offer payment b to incentivize users to sample arm $\hat{a}^*$.*

3) Self-Sustaining Phase: *For $t \in (\tau_2, T]$, the service provider stops offering payments and lets users sample arms based on their own preferences.*

We note that the parameter p plays an important role in the UCB-List: If p is too large, it takes a long time to remove arms from set U, which prolongs the exploration phase and the incentivized duration. If p is too small, then the exploration may not be long enough to ensure a logarithmic regret.

Theorem 3. (UCB-List) *Given an incentive sensitivity function $G(\cdot)$ and a fixed payment b satisfying $G(b) > 1$, if the feedback function satisfies $F(x) = \Omega(x^\alpha)$ and $F(x) = O(x^\alpha \ln x)$ for $\alpha > 1$, then the expected regret $\mathbb{E}[R_T]$ of the UCB-List policy is upper bounded by $\mathbb{E}[R_T] \leq \sum_{a \neq a^*,\, a \in A} (\frac{8}{\Delta_a} + \frac{\gamma_p \Delta_a}{G(b)-1}) \log T + O(1) + O(T^{-3})$, with a constant γ_p dependent on the parameter p. The expected accumulative payment $\mathbb{E}[B_T]$ is upper bounded by $\mathbb{E}[B_T] \leq \frac{2\gamma_p G(b)}{G(b)-1} \log T$.*

Proof Sketch. The expected time needed for initialization can be proved upper bounded by $O(1)$ trivially. Then by the law of total expectation, the expected regret up to time T can be decomposed as:

$$\mathbb{E}[R_T] \leq \underbrace{\mathbb{E}[R_{\tau_1}]}_{\textbf{(a)}} + \underbrace{\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]}_{\textbf{(b)}} + \underbrace{\mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*]}_{\textbf{(c)}} + \underbrace{T \cdot \mathbb{P}(\hat{a}^* \neq a^*)}_{\textbf{(d)}}.$$

In what follows, we will bound the four terms on the right-hand-side one by one.

(a) In the exploration phase, since the regret results from the pulls of sub-optimal arms, the expected regret at time step τ_1 can be written as $\mathbb{E}[R_{\tau_1}] = \sum_{a \neq a^*,\, a \in A} \Delta_a \mathbb{E}[T_a(\tau_1)]$, where Δ_a is defined as $\Delta_a = \mu_{a^*} - \mu_a$ for arm $a \in A$. Thus, term (a) can be bounded if we upper bound $\mathbb{E}[T_a(\tau_1)]$ for each $a \in A$. The proof is similar to that of the standard UCB policy except that we consider in addition the case where user behaviors in our model are stochastic. Thus, there exists a positive probability for users to pull arms that are not incentivized. This possibility causes regret that can be bounded based on the convergence of the feedback function. Thus, for each arm $a \in A$, we obtain $\mathbb{E}[T_a(\tau_1)] \leq \frac{8 \ln T}{\Delta_a^2} + O(1)$. By using the Chernoff-Hoeffding bound, $\mathbb{E}[\tau_1]$ can be bounded by $O(\log T)$.

(b) In the exploitation phase, the expected regret $\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]$ is upper bounded by $O(\mathbb{E}[\tau_2])$. Thus, following the definition of τ_2, we obtain the bound $O(\log T)$ for this term. Also, since the payment is stopped after the exploitation phase, the expected accumulative payment $\mathbb{E}[B_T]$ can be bounded by $b\mathbb{E}[\tau_2]$ given time τ_2. Note that the choice of τ_2 has the same analysis as in the proof of At-Least-n Greedy.

(c) The third term represents the expected regret from time τ_2 to T. Similar to the proof of Theorem 2, the expected regret during this phase is resulted from the pulling of any sub-optimal arms, and can be bounded by $O(1)$ based on the result in choosing τ_2 and some standard expectation calculations.

(d) The probability $\mathbb{P}(\hat{a}^* \neq a^*)$ can be bounded by $O(T^{-4})$ using the Chernoff-Hoeffding bound.

Combining steps **(a)**–**(d)** yields the result stated in the theorem and the proof is complete. □

5 SIMULATIONS

We conduct simulations to evaluate the performances of the At-Least-n Greedy and UCB-List policies.

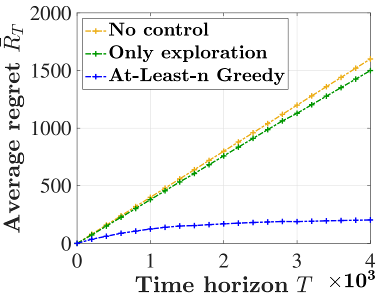
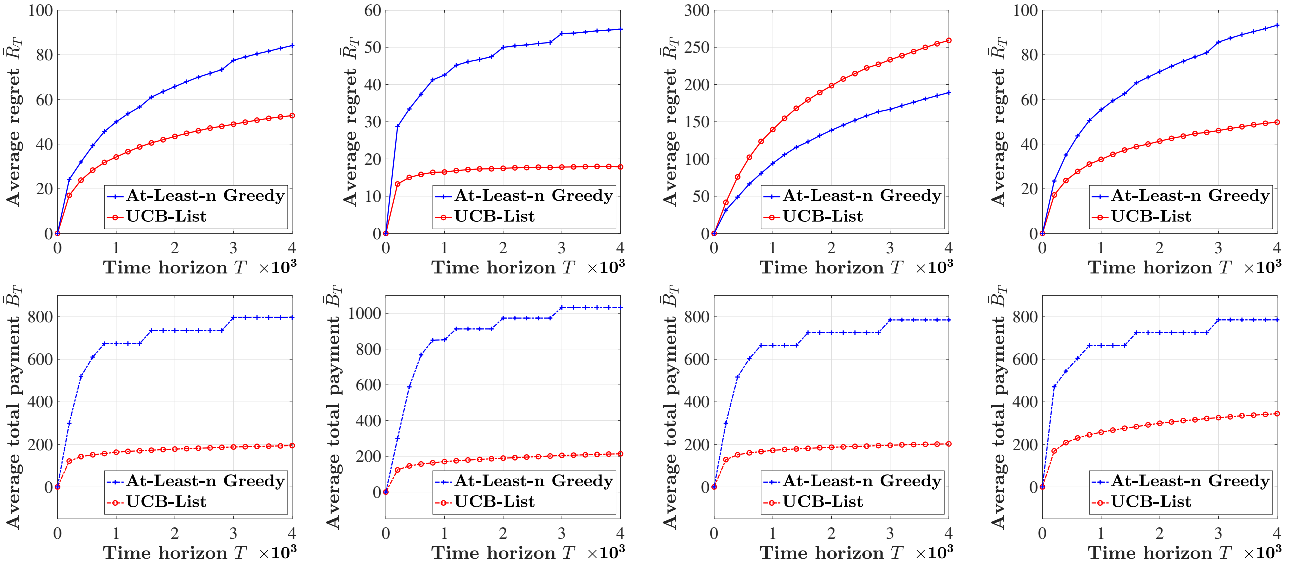


Figure 1: The performance of the At-Least-n Greedy, compared to i) the performance of the baseline with no incentive control and ii) the baseline with only incentive control during exploration phase.

1) Baseline Comparisons: We first compare our policies' performance with two baselines: i) no incentive control and ii) with incentive control only during exploration. We use At-Least-n Greedy as an example. Consider a two-armed model with means $\boldsymbol{\mu} = [0.2, 0.6]$ and initial biases $\boldsymbol{\theta} = [100, 1]$. We choose the feedback function as $F(x) = x^\alpha$ with $\alpha = 1.5$ and set the payment as $b = 1.5$ with an incentive sensitivity function $G(x) = x$. The results are illustrated in Figure 1, where each point is averaged over $10,000$ trials. We can observe that the average regret under no incentive control grows linearly due to the self-reinforcing preference on the suboptimal arm with large initial bias and large α. The average regret under partial incentive control is also linear since the incentive is insufficient to offset the initial bias from the suboptimal arm. The average regret of the At-Least-n Greedy follows a



(a) Benchmark result.

(b) $\alpha = 1.5$.

(c) $\boldsymbol{\theta} = [50, 1]$.

(d) $b = 3$.

Figure 2: Performances of the At-Least-n Greedy and UCB-List policies.

$\log(T)$ growth rate with $O(\log T)$ total payment.

2) At-Least-n Greedy vs. UCB-List: Next, we compare the performances of At-Least-n Greedy and UCB-List. The setup is the same as that in baseline comparisons, except that the initial bias is set as $\boldsymbol{\theta} = [10, 1]$ and $\alpha = 1.1$. For UCB-List, we set the parameter $p = 0.33$. As the means of arms get closer, it is better to choose a small p, since a smaller p prevents the exploration phase being too large. For At-Least-n Greedy, we set the parameter $q = 1.5$. Four groups of simulations are conducted and the results are shown in Figure 2. Figure 2a illustrates the performance of both average regret and average total payment. Figure 2a also serves as a benchmark result, which is compared to the other three groups of results. In each of the three Figures 2b–2d, only one parameter is changed compared to the benchmark group. This helps to observe the changes in average regret and average total payment. In Figure 2b, all settings are the same as Figure 2a except that $\alpha = 1.5$. In Figure 2c, all settings are the same as those in Figure 2a except that $\boldsymbol{\theta} = [50, 1]$. In Figure 2d, all settings are the same as Figure 2a except that $b = 3$.

We can see that both policies achieve $O(\log T)$ average regrets and $O(\log T)$ average total payments. This indicates that: i) both policies balance the trade-off between exploration and exploitation so that an order-optimal regret can be reached; ii) both policies balance the trade-off between maximizing the total reward and keeping the total payment grow at rate $O(\log T)$. In Figure 2b, the result shows that both policies achieve a smaller average regret, because the self-reinforcing preferences are easier to converge to the arm incentivized by the service provider under a larger α. Also, the At-Least-n Greedy policy incurs a higher total payment because it incentivizes the pulling of sub-optimal arms more often. In Figure 2c, both policies have larger average regrets because it takes more effort to mitigate the larger initial biases. Also, the average regret shows that UCB-List is more sensitive to the initial bias than At-Least-n Greedy, since UCB-List tends to make biased decisions during the early time steps. Thus, a good choice of the parameter p is important to achieve good performance under UCB-List. In Figure 2d, as the payment for each time step increases from 1.5 to 3, the average regret and average total payment are not affected significantly since the total incentivized duration decreases correspondingly.

6 CONCLUSION

We proposed and studied an incentivized bandit model with self-reinforcing preferences. Two incentivized bandit policies are proposed to achieve $O(\log T)$ expected regrets with $O(\log T)$ incentivized costs, under the condition that the feedback function satisfies $F(x) = \Omega(x^\alpha)$ and $F(x) = O(x^\alpha \ln x)$ for $\alpha > 1$. We conjecture that the feedback can be extended to a larger class of nonlinear functions. We note that the area of incentivized MAB with self-reinforcing preferences remains under-explored. Future works include, for example, incentive costs could be time-varying in each time step, which can either be dependent on the current state, or restricted by certain conditions. The self-reinforcing preferences can also be viewed as contexts, and thus this setting can be modeled by leveraging contextual bandit with more interesting properties.

References

Acemoglu, D., Dahleh, M. A., Lobel, I., and Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236.

Athreya, K. B. and Karlin, S. (1968). Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *The Annals of Mathematical Statistics*, 39(6):1801–1817.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Bawa, K. and Shoemaker, R. W. (1987). The effects of a direct mail coupon on brand choice behavior. *Journal of Marketing Research*, 24(4):370–376.

Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026.

Bouneffouf, D. and Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Chakrabarti, S., Frieze, A., and Vera, J. (2005). The influence of search engines on preferential attachment. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 293–300. Society for Industrial and Applied Mathematics.

Combes, R., Jiang, C., and Srikant, R. (2015). Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):245–257.

Davis, B. (1990). Reinforced random walk. *Probability Theory and Related Fields*, 84(2):203–229.

Drinea, E., Frieze, A., and Mitzenmacher, M. (2002). Balls and bins models with feedback. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 308–315. Society for Industrial and Applied Mathematics.

Fiez, T., Sekar, S., and Ratliff, L. J. (2018). Multi-armed bandits for correlated markovian environments with smoothed reward feedback. *arXiv preprint arXiv:1803.04008*.

Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. (2014). Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22.

Goel, A., Khanna, S., and Null, B. (2009). The ratio index for budgeted learning, with applications. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 18–27. SIAM.

Guadagni, P. M. and Little, J. D. (2008). A logit model of brand choice calibrated on scanner data. *Marketing Science*, 27(1):29–48.

Guha, S. and Munagala, K. (2007). Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 104–113.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing research*, 25(4):342–355.

Khanin, K. and Khanin, R. (2001). A probabilistic model for the establishment of neuron polarity. *Journal of Mathematical Biology*, 42(1):26–40.

Kremer, I., Mansour, Y., and Perry, M. (2014). Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Mansour, Y., Slivkins, A., and Syrgkanis, V. (2015). Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582.

Mansour, Y., Slivkins, A., Syrgkanis, V., and Wu, Z. S. (2016). Bayesian exploration: Incentivizing exploration in bayesian games. *arXiv preprint arXiv:1602.07570*.

Oliveira, R. I. (2009). The onset of dominance in balls-in-bins processes with feedback. *Random Structures & Algorithms*, 34(4):454–477.

Papatla, P. and Krishnamurthi, L. (1996). Measuring the dynamic effects of promotions on brand choice. *Journal of Marketing Research*, 33(1):20–35.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. (2010). Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15):158701.

Shah, V., Blanchet, J., and Johari, R. (2018). Bandit learning with positive externalities. In *Advances in Neural Information Processing Systems*, pages 4918–4928.

Smith, L. and Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, 68(2):371–398.

Wang, S. and Huang, L. (2018). Multi-armed bandits with compensation. In *Advances in Neural Information Processing Systems*, pages 5114–5122.

Xia, Y., Li, H., Qin, T., Yu, N., and Liu, T.-Y. (2015). Thompson sampling for budgeted multi-armed bandits. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zhu, T. (2009). Nonlinear pólya urn models and self-organizing processes. *Unpublished dissertation, University of Pennsylvania, Philadelphia*.