Incentivized Bandit Learning with Self-Reinforcing User Preferences

Tianchen Zhou [1] **Jia Liu** [1] **Chaosheng Dong** [2] **Jingyuan Deng** [2]

Abstract

In this paper, we investigate a new multi-armed bandit (MAB) online learning model that considers real-world phenomena in many recommender systems: (i) the learning agent cannot pull the arms by itself and thus has to offer payments 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. Besides addressing the tradeoff of exploration and exploitation, another key feature of this new MAB model is to balance reward and incentivizing payment. The goal of the agent is to minimize the accumulative regret over a fixed time horizon T with a low total payment. Our contributions in this paper are two-fold: (i) We propose a new MAB model with random arm selection that considers the relationship of users' self-reinforcing preferences and incentives; and (ii) We leverage the properties of a multi-color Pólya urn with nonlinear feedback models to propose two MAB policies termed "At-Least-n Explore-Then-Commit" and "UCB-List." We prove that both policies achieve $O(\log T)$ expected regret with $O(\log T)$ expected payment over a time horizon T. We conduct numerical simulations to demonstrate and verify the performances of these two policies and study their robustness under various settings.

1. Introduction

In many online e-Commerce platforms, there exists a self-reinforcing phenomenon, where the current user's behavior is influenced by the user behaviors in the past (Barabási & 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. As an online learner, the e-Commerce service provider wants to identify the most profitable item in order to maximize the total profit in the long run. In the literature, such an online profit maximization problem can often be modeled by the multi-armed bandit (MAB) framework (Berry & Fristedt, 1985; Bubeck & Cesa-Bianchi, 2012). However, existing works on MAB that consider the self-reinforcing preferences remain quite limited (see, e.g., Fiez et al. (2018); Shah et al. (2018)). In fact, Shah et al. (2018) showed that the self-reinforcing preferences might render the classic UCB (upper confidence bound) policy (Auer et al., 2002) sub-optimal, and new optimal arm selection algorithms are necessary.

On the other hand, in many online learning problems that utilize the MAB framework for sequential decision making (e.g., recommender systems, healthcare, finance, dynamic pricing, see Bouneffouf & Rish (2019)), the learning agent (e.g., an online service provider) *cannot* select the arms directly. Rather, arms are pulled by the users who are exhibiting self-reinforcing preferences. The agent 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. Meanwhile, there exist several works (Frazier et al., 2014; Mansour et al., 2015; 2016; Wang & 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.

The missing of joint modeling of incentives and self-reinforcing preferences in the existing MAB framework (two key features of many online e-Commerce systems) motivates us to fill this gap in this paper. Specifically, in this work, we first propose a more general MAB model with *stochastic arm selections following user preferences*, which is closely modeling random user behaviors in most online recommender systems. This is in stark contrast to most existing works in the areas of incentivized bandits (Frazier et al., 2014; Wang & Huang, 2018), where a (unrealistic) deterministic greedy user behavior is often assumed. Un-

[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA [2]Amazon, Seattle, Washington, USA. Correspondence to: Tianchen Zhou <zhou.2220@osu.edu>, Jia Liu <liu@ece.osu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

der this model, a pair of fundamental trade-offs naturally emerge: (1) Sufficient exploration is required to identify an optimal arm, which may result in multiple pullings of 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; (2) The agent needs to provide enough incentives to mitigate unfavorable initial bias and self-reinforcing user preferences, while in the meantime avoiding unnecessarily high incentives for users. As in most online learning problems, we use regret as a benchmark to evaluate the performance of our MAB policy, which is defined as the performance gap between the proposed policy and an optimal policy in hindsight. The major challenges in this new MAB model thus lie in the following fundamental questions:

(a) During incentivized pulling, how could the agent maintain a good balance between exploration and exploitation to minimize regret?

(b) How long should the agent incentivize until the right self-reinforcing user preference is established toward an optimal arm (so that no further incentive is needed)?

(c) Is the established self-reinforcing user preferences sufficiently strong and stable 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 answer the above questions by proposing two "$\log(T)$-regret-with-$\log(T)$-payment" policies for the incentivized MAB framework with self-reinforcing preferences. Our contributions are summarized as follows:

- We first show that no incentivized bandit policy can achieve a sub-linear regret with a sub-linear total payment if the feedback function that models the self-reinforcing preferences has a super-polynomial growth rate. 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.
- To address the unique challenges in the new MAB model, we introduce (i) a *three-phase MAB policy architecture* and (ii) a key result that shows that an $O(\log T)$ incentivizing period is sufficient for establishing *dominance* for the multi-color Pólya urn model (see Section 4). All of these results are new in the bandit literature, which could be of independent interest for other incentivized MAB problems.
- We propose two bandit policies, namely At-Least-n Explore-Then-Commit 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 & Robbins (1985). Meanwhile, the expected total incentives for both policies are upper bounded by $O(\log T)$.

2. Related Work

The self-reinforcing phenomenon has received increasing interest in several different fields recently under different terminologies. In the random network literature, previous works have studied the network evolution with "preferential attachment" (Barabási & 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 & 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 MAB algorithms to maximize the total reward. Then, Fiez et al. (2018) provided a more general model, where the learning agent has limited information. We note that the agents in Shah et al. (2018); Fiez et al. (2018) have full control in determining which arm for users to pull. In contrast, the agent in our MAB model has *no control* over which arm to pull, and can only incentivize users to indirectly induce the preferences toward a desired arm. Eventually, 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 & Huang, 2018). To our knowledge, Frazier et al. (2014) first adopted incentive schemes into a Bayesian MAB setting. In their model, the agent seeks to maximize time-discounted total reward 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 agent 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 & Huang (2018) considered the non-Bayesian setting with non-discounted rewards. Agrawal & Tulabandhula (2020) considered incentivizing exploration under contextual bandits. These models differ from ours in both the incentive schemes and user behaviors.

Another line of research similar to incentivized bandit is bandit with budgets (Guha & Munagala, 2007; Goel et al., 2009; Combes et al., 2015; Xia et al., 2015), where the agent takes actions with budget constraints. Guha & 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 agents 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 part of our goals.

Although not cast in the MAB framework, the works on *urn models* (Khanin & 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 & 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. System Model and Problem Statement

In this paper, we denote the set of arms offered by the agent as $A = \{1, \dots, m\}$. Each arm a follows a Bernoulli reward distribution D_a with an unknown mean $\mu_a > 0$. The process runs for T rounds. As shown in Fig. 1, in each time step $t \in \{1, \dots, T\}$, a user arrives and chooses an arm $I(t)$ to pull, then receives a random reward $X(t) \sim D_{I(t)}$, which is observable to the agent. We use $T_a(t) \triangleq \sum_{i=1}^{t} 1_{\{I(i)=a\}}$ to denote the number of times that an arm a is pulled up to time t. We denote the total reward generated by arm a up to time t as $S_a(t) \triangleq \sum_{i=1}^{t} X(i) \cdot 1_{\{I(i)=a\}}$. We let $T_a(0) = 0$ and $S_a(0) = 0, \forall a \in A$. We assume that there is a unique best arm $a^* \in A$, i.e., $a^* = \arg\max_a \mu_a$ and $\mu^* = \mu_{a^*}$.

1) Preference and Bias Modeling: Unlike most of the incentivized MAB models where users are rational and independent, the user behavior is *stochastic* and *influenced by history* in our model. Specifically, in each time step t, the user has a non-zero probability $\lambda_a(t) \in (0, 1)$ 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* of arm a in time step t. We adopt the widely used multinomial logit model in the literature to model $\lambda_a(t)$ as follows:

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

where $F(\cdot) : \mathbb{R} \to (0, +\infty)$ is a feedback function that is increasing, and $\theta_a > 0$ denotes the fixed initial preference bias of arm a. Intuitively, the increasing feedback function $F(\cdot)$ models the *self-reinforcing user preference effect* in the following sense: if an arm a has been more profitable in the past, a user who prefers arm a is more likely to arrive in the next round. A simple example of the feedback function is $F(x) = x^\alpha$ for some constant $\alpha > 1$. Here, α represents the strength of the self-reinforcing preference: a larger α implies a stronger self-reinforcing preference effect.

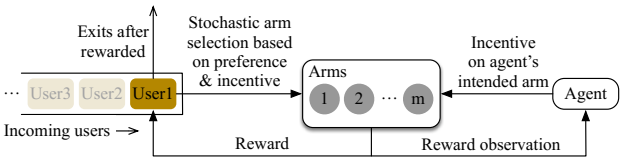


Figure 1. Incentivized MAB model with stochastic arm selection based on user preference rates and incentives.

Several important remarks for the preference model in (1) are in order. 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 & Little, 2008; Gupta, 1988). The multinomial logit model is also used in the social network literature to model preferential attachment (Barabási & 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. Notably, this multinomial logit model has also been adopted in Shah et al. (2018) to model the same type of self-reinforcing phenomenon in their MAB model.

2) Incentive Mechanism Modeling: Unlike in conventional MAB models, the agent in our model can only offer some *incentive* on the arm that the agent wants to explore, so as to increase the users' preferences of pulling this particular arm for the agent (as shown in Fig. 1). The agent's goal is to maximize total reward in the long run. In this paper, we model the influence of the incentives by adopting the so-called "coupon effects on brand choice behaviors" in the economics literature (Papatla & Krishnamurthi, 1996; Bawa & Shoemaker, 1987). In this model, the relationship between coupons and choices is nonlinear, and the redemption rate increases with respect to the coupon value but exhibits a diminishing return effect (Bawa & Shoemaker, 1987). Specifically, in time step t, if the agent wants to explore arm a, the agent will offer a fixed payment b[1] to the current user to increase the user's preference on pulling arm a. Under the coupon effect model, the posterior preference rates of the arms with incentive b are updated as follows:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{\bar{G}(b,t) + F\big(S_i(t-1) + \theta_i\big)}{\bar{G}(b,t) + \sum_{j \in A} F\big(S_j(t-1) + \theta_j\big)}, & i = a, \\ \dfrac{F\big(S_i(t-1) + \theta_i\big)}{\bar{G}(b,t) + \sum_{j \in A} F\big(S_j(t-1) + \theta_j\big)}, & i \neq a, \end{cases} \tag{2}$$

[1] In this paper, we consider fixed payment with the goal of gaining a first fundamental understanding of the regret of the proposed new MAB model. The problem of optimizing the total cost of a time-varying payment strategy is an important related problem, which will left for our future studies.

where $\bar{G}: \mathbb{R}^2 \to \mathbb{R}^+$ is an increasing function of b with $\bar{G}(0,\cdot) = 0$, which can be interpreted as the impact of payment b on users at time t. Intuitively, $\bar{G}(b,t)$ represents the "impact" of offering incentive b on users at time t. Also, $\bar{G}(b,t)$ has the property that it is increasing over time. The interpretation is that, as arms gain higher accumulative total reward $\sum_{i\in A} F\big(S_i(t-1)+\theta_i\big)$ as t increases (e.g., items gaining more positive reviews), offering the same amount of incentive b on any of them becomes more attractive.

Clearly, the posterior preference update in (2) still follows the multinomial logit model. Also, we can see from (2) that, as parameter b increases asymptotically ($b \uparrow \infty$), 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 in our subsequent 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 $G(b,t) \triangleq \bar{G}(b,t)/\sum_{i\in A} F\big(S_i(t-1)+\theta_i\big)$. Then, it can be verified that Eq. (2) can be equivalently rewritten as:

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

Clearly, $G(b,t)$ remains an increasing function of b. Also, we define the accumulative payment up to time step t as $B_t := \sum_{i=1}^{t} b_t$, where $b_t \in \{0, b\}$, $\forall t$, denotes the agent's binary decision whether to offer incentive b at time step t.

3) Regret Modeling: Let $\Gamma_T = \sum_{t=1}^{T} X(t)$ denote the accumulative reward 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 in terms of growth rate with respect to 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 conventional MAB problems, we measure our accumulative reward performance against an oracle policy, where in hindsight the agent 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 $\mathbb{E}[\Gamma_T^*] = \mu_{a^*}T$.[2] The expected (pseudo) regret is defined as: $\mathbb{E}[R_T] = \mu_{a^*}T - \mathbb{E}[\Gamma_T]$. Our goal is to minimize $\mathbb{E}[R_T]$, with low expected accumulative payment $\mathbb{E}[B_T]$ with respect to the time horizon T.

[2] 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 agent 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 agent has to spend time mitigating these initial biases, resulting in an expected accumulative reward smaller than $\mu_{a^*}T$.

4. Policy Designs and Performance Analysis

In this section, we present two policies that achieve $O(\log T)$ expected regret with $O(\log T)$ accumulative payment with respect to time horizon T.

4.1. The Basic Idea

The main idea of our two proposed policies is based on a unique *three-phase MAB policy architecture*: 1) 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; 2) We keep incentivizing the pulling of the best-empirical arm $\hat{a}^*$ until it dominates and attracts users who favor this arm; and 3) We stop incentivizing and rely on the self-reinforcing user preference to continue pulling the optimal arm. The success of our incentivized policy designs relies on guaranteeing the *dominance* of arm $\hat{a}^*$, which is defined as follows:

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

Our MAB policy designs are based on a *key fact* that, if the feedback function $F(x)$'s growth rate is superlinear polynomial, then as soon as dominance is established, we can stop incentivizing and rely on the users' self-reinforcing preferences to converge to one arm within a finite number of rounds, i.e., an arm $a \in A$ is the only arm to be sampled eventually. We call this event as the *monopoly by arm* a ($mono_a$ for short). We point out that a **key contribution** in this paper is the insight that dominance happens *much sooner than* establishing monopoly (to be shown later that this only takes $O(\log(T))$ rounds). This fact further implies the existence of an incentivized policy with *sub-linear* total payment. We formally state this fact as follows:

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* $\sum_{i=1}^{+\infty}\big(1/F(i)\big) < +\infty$.

Proof Sketch of Lemma 1. Our main technique for proving Lemma 1 is an improved exponential embedding method. This method simulates the reward generating sequence by random exponentials. In what follows, we outline the key steps of the proof and relegate the details to the supplementary material.

In contrast, we assume that the agent 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. As a result, 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.

Step 1) Construction of an Equivalent Reward Generating Sequence: 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 if 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 $\mathcal{F}_{j-1}$. Then, the constructed sequence $\{\zeta_j\}$ will be leveraged to prove the lemma.

For arm i, consider a collection of independent exponential random variables $\{r_i(n)\}$ such that $\mathbb{E}[r_i(n)] = 1/[\mu_i F(n+\theta_i)]$. 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 obtain accumulative reward n. Then we mix and sort B_i in an increasing order for all $i \in A$ to form a new sequence H. Our objective sequence $\{\zeta_j\}$ is the arm index sequence out of H. Then, we can prove by induction that given the previous reward history $\mathcal{F}_{j-1}$, the constructed sequence $\{\zeta_j\}$ has the same conditional distribution as $\{\chi_j\}$.

Step 2) Establishing Attraction Time: The proof of Lemma 1 is done once we show that if and only if any feedback function $F(x) > 0$ satisfies $\sum_i \big(1/F(i)\big) < +\infty$, then $\mathbb{P}(\exists a \in A,\, mono_a) = 1$. We define the *attraction time* N as the time step when the monopoly happens. With the constructed sequence $\{\zeta_j\}$, we establish the necessity by showing that if $\sum_i \big(1/F(i)\big) < +\infty$ then $\mathbb{P}(N < \infty) = 1$, and the sufficiency by showing that if $\sum_i \big(1/F(i)\big) = +\infty$ then $\mathbb{P}(N = \infty) > 0$. This completes the proof. □

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

Remark 2. A simple example that satisfies the condition in Lemma 1 is $F(x) = Cx^{\alpha}$ for some constants $C > 0$ and $\alpha > 1$ (i.e., superlinear polynomial). In this case, there exists an incentivized policy that induces all preferences to converge over time with sub-linear total payment, since $\sum_{i=1}^{+\infty}(1/i^{\alpha}) < +\infty$ with $\alpha > 1$. Previous works (Drinea et al., 2002; Khanin & Khanin, 2001) considering the balls and bins model also studied this feedback function with $\alpha \leq 1$. For $\alpha < 1$, the asymptotic preference rates of arms are all deterministic, positive, and dependent on the means and biases of arms. For $\alpha = 1$, the system is akin to a standard Pólya urn model, and will converge to a state where all arms have random positive preference rates depending on the means and initial biases of the arms. For $\alpha > 1$, 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, systems under these three α-values exhibit completely different behaviors.

Remark 3. In our later theoretical and numerical studies in this paper, we will focus on the class of polynomial functions $F(x) = \Theta(x^{\alpha})$ with $\alpha > 1$ as the feedback function. We note that the use of $F(x) = \Theta(x^{\alpha})$ does not lose much generality since all analytic functions in a bounded range can be approximated arbitrarily well by their Taylor polynomial expansions. Also, since $F(x)$ that satisfies the condition $\sum_{i=1}^{+\infty}\big(1/F(i)\big) < +\infty$ in Lemma 1 is lower bounded by $\Omega(x^{\alpha})$ with $\alpha > 1$ (by considering $\sum_{i=1}^{+\infty}\big(1/F(i)\big)$ as p-series), $F(x) = \Theta(x^{\alpha})$ with $\alpha > 1$ is general enough to cover a large class of functions.

4.2. The At-Least-n Explore-Then-Commit Policy

Our first policy is the At-Least-n Explore-Then-Commit (ALnETC), which consists of three phases: the exploration phase, the exploitation phase, and the self-sustaining phase. The agent incentivizes in the first two phases. During the exploration phase, ALnETC explores all arms until each arm generates sufficient accumulative reward. Then, the policy incentivizes the arm with the best empirical mean until it *dominates* (as defined in Definition 1). Toward this end, we define the sample mean of arm $a \in A$ at time step t as $\hat{\mu}_a(t) = S_a(t-1)/T_a(t-1)$. Then, we formally state the ALnETC policy as follows:

Policy 1: At-Least-n Explore-Then-Commit

Given time horizon T, payment b and $n = q \ln T$, where $q > 0$ is some tuning parameter:

1) Exploration Phase: Incentivize pulling arm $a \in \arg\min_{i \in A} S_i(t)$ with payment b until time $\tau_n = \min\{t : S_a(t) \geq n, \forall a\} \wedge T$, when any arm has accumulative reward of at least n.

2) Exploitation Phase: Incentivize pulling the best-empirical arm $\hat{a}^* \in \arg\max_{a \in A} \hat{\mu}_a(\tau_n)$ with payment b until it dominates, i.e., $S_{\hat{a}^*}(t) \geq \sum_{a \neq \hat{a}^*} S_a(t)$. Mark current time as τ_s.

3) Self-Sustaining Phase: Users pull arms based on their own preferences until time T.

For the ALnETC policy, we next show that if the incentive effect is sufficiently strong, then the dominance time τ_s happens within $O(\log T)$ rounds, which is much sooner than the attraction time (i.e., time for establishing monopoly). We formally state this result as follows:

Lemma 2. (Dominance) *In ALnETC, if the incentive sensitivity function $G(\cdot)$ and the payment b satisfy $G(b,t) > 1$ for all t in the exploration and exploitation phases, then the*

expected dominant time τ_s is $O(\log T)$.

Remark 4. In Lemma 2, the condition "$G(b,t) > 1$" has an interesting interpretation in practice. Recall that $G(b,t)$ is defined as $G(b,t) \triangleq \bar{G}(b,t)/\sum_{i\in A} F\big(S_i(t-1)+\theta_i\big)$ (cf. Section 3). Thus, $G(b,t) > 1$ means that the "incentive impact" $\bar{G}(b,t)$ should be larger (could be ever so slightly) than the "impact of arms' accumulative reward" $\sum_{i\in A} F(S_i(t-1)+\theta_i)$ so that incentive control is possible.

Based on the above result, we will show next that once the best-empirical arm *dominates*, then it implies sub-linear regret and accumulative incentive payment. Intuitively, this is because we will show that, within a finite number of steps after dominance time τ_s, monopoly happens with probability one, and arm $\hat{a}^*$ has a high probability to emerge victorious in the monopoly (to be shown in the proof of Theorem 3). If the time horizon T is sufficiently large to cover the attraction time (i.e., the time when monopoly happens), then arm $\hat{a}^*$ will be sampled repeatedly after the attraction time, while the expected pulling times from sub-optimal empirical arms after the dominance is $o(\log T)$ (which contributes to the regret). Thus, the policy achieves a sub-linear expected regret. For each arm a, we set $\Delta_a = \mu^* - \mu_a$, and let $\Delta_{min} = \min_{a\neq a^*} \Delta_a$, $\Delta_{max} = \max_{a\neq a^*} \Delta_a$. We formally state this result as follows:

Theorem 3. (At-Least-n Explore-Then-Commit) *Given a fixed time horizon T, if (i) $G(b,t) > 1$, (ii) $q \geq (2\max_{a\neq a^*}\mu_a)/\Delta^2_{min}$, (iii) $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then the expected regret of ALnETC is upper bounded by:*

$$\mathbb{E}[R_T] \leq \sum_{a\in A} \frac{2(G(b,t)-L_{a^*})\Delta_{max}}{\big(G(b,t)-1\big)\mu_a} \cdot q\ln T + o(\log T),$$

where $L_a = F(q\ln T + \theta_a)/\sum_{i\in A} F(\mu^* T + \theta_i)$. The expected total payment is upper bounded by:

$$\mathbb{E}[B_T] \leq \sum_{a\neq a^*} \frac{2b(G(b,t)+1)}{\mu_a(G(b,t)-1)} \cdot q\ln T.$$

Remark 5. For a given incentive b, as $G(b,t)$ increases asymptotically (large incentive impact), regret and total payment decrease to some limiting amounts. This makes intuitive sense since if the incentive has a larger impact on users, it will reduce the pullings of random unfavorable arms and shorten the exploration and exploitation phases. On the other hand, as $G(b,t)$ decreases towards one from above, users are less affected by incentives, thus in many instances the exploration phase never stops. This could lead to linear expected regret and linear expected total payment. Meanwhile, as q decreases, both regret and total payment are smaller. But if $q < (2\max_{a\neq a^*}\mu_a)/\Delta^2_{min}$, the exploration will be insufficient to guarantee the event $\{\hat{a}^* = a^*\}$. This leads to a linear regret. Also, a large Δ_{max} implies larger a loss of pullings of suboptimal arms to reach n accumulative reward during exploration phase, leading to a larger regret.

Proof Sketch of Theorem 3. Due to space limitation, we provide a proof sketch here and relegate the details to the supplementary material. By the law of total expectation, the expected regret up to time T can be decomposed as:

$$\mathbb{E}[R_T] \leq \underbrace{\mathbb{E}[R_T \mid \hat{a}^* = a^*]}_{\text{(a)}} + T \cdot \underbrace{\mathbb{P}(\hat{a}^* \neq a^*)}_{\text{(b)}}.$$

To bound $\mathbb{E}[R_T]$, we want to upper bound both $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$ and $\mathbb{P}(\hat{a}^* \neq a^*)$. First, in (b), the probability $\mathbb{P}(\hat{a}^* = a^*) \leq \mathbb{P}\big(\hat{\mu}_a(\tau_n) \geq \hat{\mu}_{a^*}(\tau_n)\big)$ is bounded by $O(T^{-1})$ by leveraging the Chernoff-Hoeffding bound. Also, noting that

$$(a) = \mu^* T - \big(\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*] + \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*]\big),$$

where Γ_t is the accumulative reward up to time t, we first need to upper 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 $\mu_a G(b,t)/\big(G(b,t)+1\big)$. Then, for each arm $a \in A$ to obtain at least n accumulative reward, the expected time needed is upper bounded by

$$\mathbb{E}[\tau_n] \leq \frac{G(b,t)+1}{G(b,t)} \cdot \sum_{i\in A} \frac{q\ln T}{\mu_i}.$$

For $\mathbb{E}[\tau_s]$, since τ_s is the earliest time for the system to reach dominance, τ_s satisfies the condition $\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a\neq\hat{a}^*} \mu_a \mathbb{E}[T_a(t)]$. With the bound of $\mathbb{E}[\tau_n]$, after relaxing the inequality and some rearrangement, we obtain the upper bound as follows:

$$\mathbb{E}[\tau_s] \leq \frac{G(b,t)+1}{G(b,t)-1} \cdot \sum_{a\neq a^*} \frac{2q\ln T}{\mu_a}.$$

According to the policy, the expected accumulative payment $\mathbb{E}[B_T]$ can be bounded by $b\mathbb{E}[\tau_s]$ and part of the expected regret $\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*]$.

The next challenge is to show whether the dominant arm has a large enough probability to "win" in monopoly during the self-sustaining phase. We use $D(u_0, n_0)$ to denote the "bad event" that the fraction of accumulative reward from weak arms increases over time. Formally, suppose that at time step τ_s, there are $u_0 n_0$ accumulative reward generated by weak arms, where n_0 is the total reward and $u_0 < 1/2$ is the fraction. Then, $D(u_0, n_0)$ happens if $\exists t' \in (\tau_s, T]$, un accumulative reward is generated from weak arms with fraction $u > u_0$. The probability of event $D(u_0, n_0)$ can be bounded as $\mathbb{P}\big(\exists n > n_0,\, D(u_0, n_0)\big) \leq e^{-(u_0 n_0)^\gamma} = e^{-O(\log T)^\gamma}$ with constant $\gamma \in (0, 1/4)$ using the improved exponential embedding method and a Chernoff-like bound developed in the supplementary material. The upper bound of event $D(u_0, n_0)$ decreases as $u_0 n_0$ increases monotonically over time. Thus, the arms that stay on the weak side for a long time have little chance to win back.

Lastly, we bound the term $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$ in (a), which contributes to the $o(\log T)$ regret term in Theorem 3. After time τ_s, a unit reward is generated by sub-optimal arms with probability upper bounded by $e^{-(u_0 n_0)^\gamma}$, and then the next unit reward is also generated by sub-optimal arms with probability upper bounded by $e^{-(u_0 n_0+1)^\gamma}$. Thus,

$$\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*] \leq e^{-(u_0 n_0)^\gamma} + e^{-(u_0 n_0+1)^\gamma} + \cdots,$$

with the summation on the right hand side bounded by $O\big((\log T)^{1-\gamma} e^{-(\log T)^\gamma}\big)$ and $\gamma \in (0, 1/4)$. □

4.3. The UCB-List Policy

In this section, we propose a UCB-List policy to further improve the performance of the ALnETC policy. UCB-List is similar to ALnETC and also consists of three phases. During the exploration phase, the agent initially puts all arms in one set, and then incentivizes the least pulled arm in the set. 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. Note that in this phase, users can still pull any arm regardless of the set. Then, the agent incentivizes users to sample the best-empirical arm until it dominates. The UCB-list policy is stated as follows:

Policy 2: The UCB-List Policy

Given time horizon T and payment b, define the confidence interval of arm a at time step t as $c_a(t) = \sqrt{\ln T / 2T_a(t)}$:

Initialization: Incentivize pulling arms satisfying $T_a(t) = 0$ with payment b until $\min_{a \in A} T_a(t) = 1$. Let set $U = A$.

1) Exploration Phase: While $|U| > 1$, keep removing any arm a satisfying $\hat{\mu}_a(t) + c_a(t) \leq \max_{i \neq a, i \in U} \big(\hat{\mu}_i(t) - c_i(t)\big)$ from U if there is any. Then, incentivize pulling arm $a \in \arg\min_{i \in U} T_i(t)$ with payment b. If $|U| = 1$, let arm $\hat{a}^* = \{a : a \in U\}$ and mark current time as τ_1.

2) Exploitation Phase: Incentivize pulling arm $\hat{a}^*$ with payment b until it dominates: $S_{\hat{a}^*}(t) \geq \sum_{a \neq \hat{a}^*} S_a(t)$. Mark current time as τ_s.

3) Self-Sustaining Phase: Users pull arms based on their own preferences until time T.

Compared to ALnETC that requires a tuning parameter q, UCB-List does not need any tuning parameter and dynamically eliminates suboptimal arms, while still balancing the exploration-exploitation trade-off to achieve $O(\log(T))$ regret and $O(\log(T))$ payment. We state this result as follows:

Theorem 4. (UCB-List) *Given a fixed time horizon T, if $G(b,t) > 1$, and $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then the expected regret of UCB-List $\mathbb{E}[R_T]$ is upper bounded by*

$$\sum_{a \neq a^*} \left[\frac{8\Delta_a \big(G(b,t)-1\big) + 8\Delta_{max}}{\big(G(b,t)-1\big)\Delta_a^2} \ln T + 4\Delta_a + \frac{4\Delta_{max}}{G(b,t)-1} \right],$$

with the expected payment $\mathbb{E}[B_T]$ upper bounded by

$$\frac{2G(b,t)+1}{G(b,t)-1} \left[\frac{8b \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \left(\frac{8b \ln T}{\Delta_a^2} + 4b \right) \right].$$

Remark 6. Without any tuning parameter, the UCB-List policy adapts to a larger range of systems. The system parameters such as means of arms $\boldsymbol{\mu}$ or their gap summation $\sum_{a \neq a^*} \Delta_a$ play an important role in both regret and total payment. As $\sum_{a \neq a^*} \Delta_a$ decreases (implying it is harder to differentiate a^*), longer exploration and exploitation phases are needed, resulting in larger expected regret and total payment. Also, similar to Theorem 3, as $G(b,t) \downarrow 1$, the expected regret and expected total payment are closer to being linear, because of the weak incentive effect.

Proof Sketch of Theorem 4. We provide a proof sketch here and relegate the details to the supplementary material. The expected time for initialization can be upper bounded by $O(1)$ trivially. By the law of total expectation, we have:

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

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 $\mathbb{E}[R_{\tau_1}] = \sum_{a \neq a^*} \Delta_a \mathbb{E}[T_a(\tau_1)]$. Thus, term **(a)** can be bounded if we upper bound $\mathbb{E}[T_a(\tau_1)]$ for each $a \in A$. Let $U(t)$ denote the set of arms that can get payment at time t. Consider the following two cases: **(i)** At time $t \leq \tau_1$, $a^* \in U(t)$ and there exists at least one suboptimal arm $a \in A, a \neq a^*$ such that $a \in U(t)$. In this case we upper bound the probability $\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \in U(t)\big)$, and by using the Chernoff-Hoeffding bound, we obtain that when $T_a(t) \geq (8 \ln T)/\Delta_a^2$ we have $\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \in U(t)\big) \leq 2T^{-1}$. Thus, in this case, the expected regret is contributed by a suboptimal arm a is $\Delta_a \mathbb{E}[T_a(t)] \leq (8 \ln T)/\Delta_a + 2\Delta_a$; **(ii)** At time $t \leq \tau_1$, a^* is eliminated by some suboptimal arm $a \in U(t)$. With the Chernoff-Hoeffding bound, we obtain $\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \notin U(t)\big) \leq 2T^{-1}$. Summing over all possible cases and all suboptimal arms, $\mathbb{E}[R_{\tau_1}]$ is bounded by:

$$\mathbb{E}[R_{\tau_1}] \leq \sum_{a \neq a^*} \frac{8 \ln T}{\Delta_a} + 4\Delta_a.$$

(b) In the exploitation phase, the expected regret $\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]$ is upper bounded by $O(\mathbb{E}[\tau_2 - \tau_1])$ since

$$\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*] \leq \frac{\Delta_{max}}{G(b)+1} \cdot \mathbb{E}[\tau_2 - \tau_1].$$

In term **(a)**, the upper bound of $\mathbb{E}[R_{\tau_1}]$ implies that each suboptimal arm a is pulled at least $(8 \ln T)/\Delta_a^2$ with a^* being pulled at least $(8 \ln T)/\Delta_{min}^2$ times, similar to the proof of Theorem 3 we obtain the upper bound of both $\mathbb{E}[\tau_1]$ and $\mathbb{E}[\tau_2 - \tau_1]$. This leads to the upper bounds of both $\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]$ and $\mathbb{E}[B_T] = (\mathbb{E}[\tau_1] + \mathbb{E}[\tau_s - \tau_1])b$.

(c) This term represents the expected regret from τ_2 to T. Similar to the proof of Theorem 3, this part of expected regret is bounded by $O\big((\log T)^{1-\gamma} e^{-(\log T)^{\gamma}}\big)$, $\gamma \in (0, 1/4)$.

(d) The probability $\mathbb{P}(\hat{a}^* \neq a^*)$ can be bounded by $O(T^{-1})$ since $\mathbb{P}(\hat{a}^* \neq a^*) = \mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \notin U(t)\big)$, which can be bounded by $2T^{-1}$ as in **(a)**-case **(ii)**.

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

5. Simulations

In this section, we conduct simulations to evaluate the performances of ALnETC and UCB-List policies.

5.1. Comparisons with Baselines

We first compare the ALnETC policy with two baselines: i) no incentive control, and ii) with incentive control only during exploration. We only compare ALnETC with the baselines since UCB-List outperforms ALnETC (to be discussed next). The simulation setting is as follows: a two-armed model with means $\boldsymbol{\mu} = [0.3, 0.5]$ and initial biases $\boldsymbol{\theta} = [100, 1]$, the feedback function $F(x) = x^{\alpha}$ with $\alpha = 1.5$ and payment $b = 1.5$ with an incentive impact function $G(x, t) = x$. We use the optimal ALnETC parameter $q = 15$. The results are shown in Fig. 2, where each data point is averaged over 1000 trials. We observe that the average regret under no incentives grows linearly due to the large initial bias toward the suboptimal arm and self-reinforcing preferences. The average regret under partial incentive is also linear since the incentive is insufficient to offset the initial bias toward the suboptimal arm. In contrast, the average regret of ALnETC policy follows a $\log(T)$ growth rate.

5.2. Comparisons with Imperfect Conditions

In real-world applications, some of our model conditions may not always hold (e.g., the conditions $G(b, t) > 1$ and $F(x) = \Theta(x^{\alpha})$ with $\alpha > 1$). Therefore, we conduct the simulations to study the robustness of our proposed policies. The system setting in the group with incentive is almost the same as that in Section 5.1: a two-armed model with means $\boldsymbol{\mu} = [0.3, 0.5]$ and initial biases $\boldsymbol{\theta} = [100, 1]$, the feedback function $F(x) = x^{\alpha}$. The key difference is that, in this study, we set $\alpha \leq 1$ and $G(b, t) < 1$ (i.e., the conditions in our theoretical results are not satisfied). Specifically, we set the value of $G(b, t)$ to be 0.5 and 0.2, implying a weaker incentive impact. Also, we choose the value of α to be 1.0 and 0.2, implying a weaker self-reinforcing preference strength. We use the optimal ALnETC parameter $q = 15$. The results are shown in Fig. 3, where each data point is averaged over 1000 trials. We observe that as the values of α and $G(b, t)$ decrease, the average regrets of both policies increase. Specifically, when the incentive impact $G(b, t)$ becomes small enough, or the self-reinforcing preference strength is weak enough (e.g., $\alpha \leq 1$), the regrets of both policies no longer exhibit sub-linear trends.

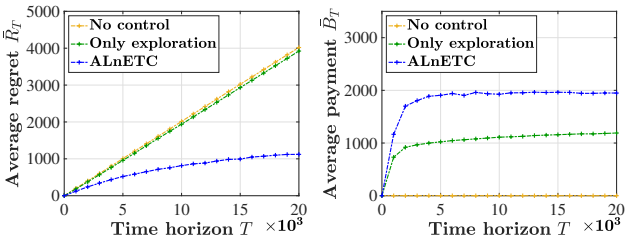
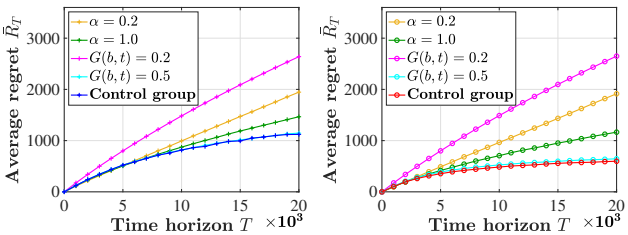


Figure 2. Comparison of ALnETC and baselines.



(a) Performance of ALnETC. (b) Performance of UCB-List.

Figure 3. Comparisons of imperfect conditions.

5.3. Comparisons between ALnETC and UCB-List

Finally, we compare ALnETC and UCB-List. The simulation setting is as follows: a three-armed model with means $\boldsymbol{\mu} = [0.2, 0.4, 0.6]$ and initial biases $\boldsymbol{\theta} = [10, 10, 1]$, the feedback function $F(x) = x^{\alpha}$, $\alpha = 1.5$ and payment $b = 1.2$ with an incentive impact function $G(x, t) = x$. For ALnETC, we set the optimal parameter $q = 20$. Four groups of simulations are conducted and the results are shown in Fig. 4-7, where each data point is averaged over 1000 trials. Fig. 4 illustrates the performance of both average regret and total payment. Fig. 4 also serves as a benchmark for comparisons with other three groups of results. In each of Figs. 5–7, only one parameter is changed compared

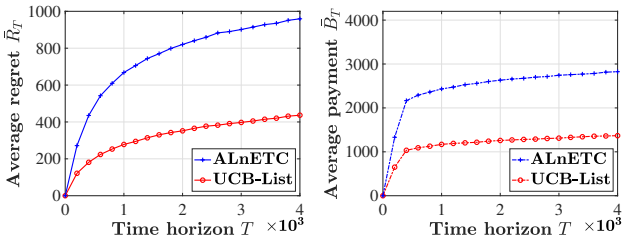


Figure 4. Benchmark results.

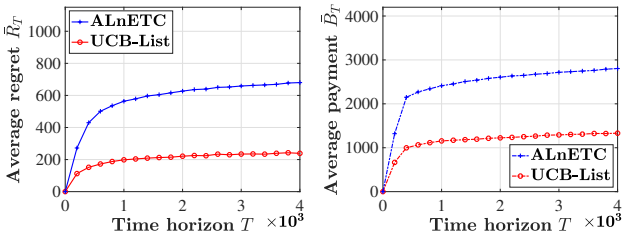


Figure 5. Policy performance with parameter $\alpha = 2$.

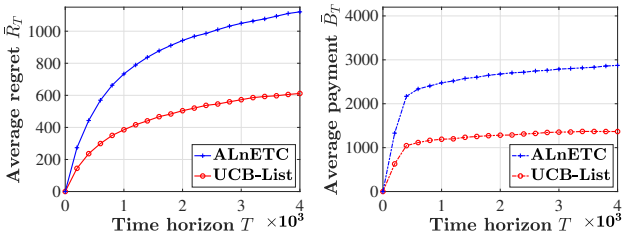


Figure 6. Policy performance with parameter $\boldsymbol{\theta} = [50, 50, 1]$.

to the benchmark group. This helps us observe the changes in average regret and total payment. In Fig. 5, all settings are the same as Fig. 4 except $\alpha = 2$. In Fig. 6, all settings are the same as those in Fig. 4 except $\boldsymbol{\theta} = [50, 50, 1]$. In Fig. 7, all settings are the same as Fig. 4 except $b = 1.8$.

The results show that both policies achieve $O(\log T)$ average regrets and $O(\log T)$ average total payment. This indicates that: i) both policies balance the exploration-exploitation trade-off 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 growing at rate $O(\log T)$. In Fig. 5, the results show that both policies achieve a smaller average regret, because the self-reinforcing preferences are easier to converge to the incentivized arm under a larger α. Also, ALnETC incurs a higher total payment because it incentivizes the pulling of sub-optimal arms more often. In Fig. 6, both policies have larger average regrets because it takes more effort for both policies to mitigate the larger initial biases. In Fig. 7, as the payment for each time step increases from 1.5 to 1.8, the average regrets are not affected significantly, while the total payments increases correspondingly. Thus, a proper amount of payment depends on specific system parameters.

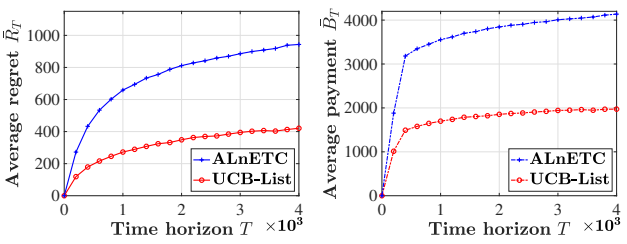


Figure 7. Policy performance with parameter $b = 1.8$.

6. Conclusion

We proposed and studied an incentivized bandit model with self-reinforcing preferences. Two policies are proposed to achieve $O(\log T)$ expected regrets with $O(\log T)$ incentivized costs, under the condition that the feedback function satisfies $F(x) = \Theta(x^\alpha)$ 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, the design of incentive schemes that can be time-varying in each time step, which can either depend on the current state, or be restricted by certain conditions. The self-reinforcing preferences can also be viewed as contexts, and thus this setting can be modeled by leveraging the contextual bandit framework with more interesting properties.

Acknowledgements

This work has been supported in part by NSF grants CAREER CNS-2110259, CCF-2110252, ONR grant N00014-17-1-2417, and a Google Faculty Research Award.

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

References

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

Agrawal, P. and Tulabandhula, T. Incentivising exploration and recommendations for contextual bandits with payments. In *Multi-Agent Systems and Agreement Technologies*, pp. 159–170. Springer, 2020.

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

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time

analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplementary Material

A. Proof of Lemma 1

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 $\sum_{i=1}^{+\infty}\big(1/F(i)\big) < +\infty$.*

Let the sequence $\{\chi_j\}_{j=1}^{\infty}$ be the arm order that generates a unit reward in our model without the participation of incentive, such that χ_j indicates the arm that generates the j-th unit reward, as shown in Figure 8. Next we will construct a sequence that has the same conditional distribution as $\{\chi_j\}$.

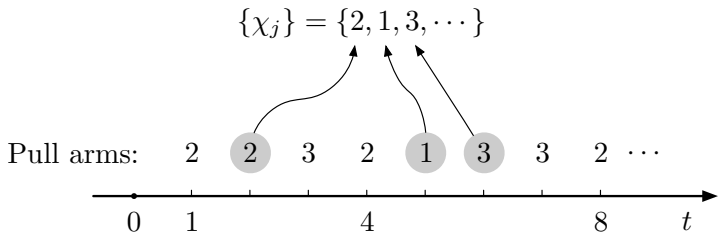


Figure 8. This figure shows an instance of sequence $\{\chi_j\}$. At time step $t = 1$, arm 2 is pulled and generates 0 reward. At time step $t = 2$, arm 2 is pulled and generates a unit reward. Thus, the first element χ_1 in $\{\chi_j\}$ is the arm index 2 that generates the first unit reward. The subsequent elements in the sequence are generated similarly.

Our main mathematical tool is the *improved exponential embedding* method. For each arm $i \in A$, we let $\{r_i(n)\}$ be a collection of independent exponential random variables such that $\mathbb{E}[r_i(n)] = \frac{1}{\mu_i F(n+\theta_i)}$. We define set $B_i := \{\sum_{k=0}^{n} r_i(k)\}_{n=0}^{\infty}$, where each element $\sum_{k=0}^{n} r_i(k)$ represents the random time needed for arm i to get n accumulative reward, and define set $G = B_1 \cup B_2 \cup \cdots \cup B_m$. Let ζ_1 be the smallest number in G and in general let ζ_j be the j-th smallest number in G. Next, we define a new random sequence $\{\zeta_j\}$, by making the j-th element of the sequence be the arm i if $\zeta_j \in B_i$. Then, we have the following lemma (to be proved later):

Lemma 5. *Given the previous reward history $\mathcal{F}_{j-1}$, the constructed sequence $\{\zeta_j\}$ is equivalent in conditional distribution to the sequence $\{\chi_j\}$.*

Next, we formally define the notion of attraction time.

Definition 2 (Attraction time)**.** *Let N denote the attraction time, such that after this time step N, monopoly happens, i.e., only one arm has positive probability to generate rewards.*

Necessity: if $\alpha > 1$ then $\mathbb{P}(N < \infty) = 1$. With the help of improved exponential embedding, the time until the accumulative reward of arm $i \in A$ approaches infinity is $\sum_{k=0}^{\infty} r_i(k)$. If the condition $\sum_i \frac{1}{F(i)} < \infty$ is satisfied, then we have

$$\mathbb{E}\big[\sum_{k=0}^{\infty} r_i(k)\big] = \frac{1}{\mu_i}\sum_{k=0}^{\infty}\frac{1}{F(k+\theta_i)} < \infty.$$

So for each arm $i \in A$, $\mathbb{P}(\sum_{k=0}^{\infty} r_i(k) < \infty) = 1$. Let $a = \arg\min_{i\in A}\{\sum_{k=0}^{\infty} r_i(k)\}$, then for each $b \neq a$, there exists a finite number K_b such that

$$\sum_{k=0}^{K_b} r_b(k) < \sum_{k=0}^{\infty} r_a(k) < \sum_{k=0}^{K_b+1} r_b(k).$$

Thus if we let $N := \max_{i\in A, i\neq a}\{\sum_{k=0}^{f_i(k)} r_i(k)\}$, then after this time N, only arm a can generate rewards.

Sufficiency: if $\mathbb{P}(N < \infty) = 1$ then $\sum_i \frac{1}{F(i)} < \infty$. If we show that when $\sum_i \frac{1}{F(i)} = \infty$ we have $\mathbb{P}(N = \infty) > 0$, then the proof is done. When $\sum_i \frac{1}{F(i)} = \infty$, we have

$$\mathbb{E}\big[\sum_{k=0}^{\infty} r_i(k)\big] = \frac{1}{\mu_i}\sum_{k=0}^{\infty}\frac{1}{F(k+\theta_i)} \to \infty.$$

Thus for any $i \in A$ it takes infinite time to accumulate infinite reward, which implies $\mathbb{P}(N = \infty) > 0$. In fact, in this case $\mathbb{P}(N = \infty) = 1$. We refer readers to Khanin & Khanin (2001) and Oliveira (2009) for further details.

A.1. Proof of Lemma 5

The proof of this lemma relies on the memoryless property of the exponential distribution as well as the following two facts:

Fact 1. *If $X_1, \cdots, X_m (m \geq 2)$ are independent exponential random variables with parameter $\lambda_1, \cdots, \lambda_m$, respectively, then $\min(X_1, \cdots, X_m)$ is also exponential with parameter $\lambda_1 + \cdots + \lambda_m$.*

Fact 2. *For two independent exponential random variables $X_1 \sim exp(\lambda_1)$ and $X_2 \sim exp(\lambda_2)$, $\mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.*

Initially, in the sequence $\{\zeta_j\}$ when $j = 1$, since the initial value for arm i is its bias θ_i, using the above two facts:

$$\begin{aligned}\mathbb{P}(\zeta_1 = i \mid \mathcal{F}_0) &= \mathbb{P}\left(r_i(0) < \min_{j \neq i}\{r_j(0)\} \middle| \mathcal{F}_0\right) \\ &= \frac{\mu_i F(\theta_i)}{\sum_{j \in A} \mu_j F(\theta_j)}.\end{aligned}$$

In our model, each arm i has probability $\mu_i \cdot \lambda_i(t) = \frac{\mu_i F(\theta_i)}{\sum_{j \in A} F(\theta_j)}$ to generate the first reward every time step before it does. The value of element χ_1 is a random variable following multinomial distribution with single trial, i.e., with $\mathcal{F}_0$, the event $\{\chi_1 = i\}$ happens with probability $\mathbb{P}(\chi_1 = i \mid \mathcal{F}_0) = \frac{\mu_i F(\theta_i)}{\sum_{j \in A} \mu_j F(\theta_j)}$, and $\sum_{i \in A} \mathbb{P}(\chi_1 = i \mid \mathcal{F}_0) = 1$. Thus

$$\mathbb{P}(\zeta_1 = i \mid \mathcal{F}_0) = \mathbb{P}(\chi_1 = i \mid \mathcal{F}_0)$$

Now suppose that before ζ_n, each arm a has been added to N_a. Then

$$\begin{aligned}\mathbb{P}(\zeta_n = i \mid \mathcal{F}_{\zeta_{n-1}}) &= \mathbb{P}\left(r_i(N_i + 1) < \min_{j \neq i}\{r_j(N_j + 1)\} \middle| \mathcal{F}_{\zeta_{n-1}}\right) \\ &= \frac{\mu_i F(N_i + \theta_i)}{\sum_{j \in A} \mu_j F(N_j + \theta_j)}.\end{aligned}$$

Correspondingly in our model, each arm i has probability $\mu_i \cdot \lambda_i(t) = \frac{\mu_i F(N_i + \theta_i)}{\sum_{j \in A} F(N_j + \theta_j)}$ to generate the next reward every time step before it does. The value of element χ_n is a random variable following multinomial distribution with single trial, i.e., with $\mathcal{F}_{\chi_{n-1}}$, the event $\{\chi_n = i\}$ happens with probability $\mathbb{P}(\chi_n = i \mid \mathcal{F}_{\chi_{n-1}}) = \frac{\mu_i F(N_i + \theta_i)}{\sum_{j \in A} \mu_j F(N_j + \theta_j)}$, and $\sum_{i \in A} \mathbb{P}(\chi_n = i \mid \mathcal{F}_{\chi_{n-1}}) = 1$. Thus,

$$\mathbb{P}(\zeta_n = i \mid \mathcal{F}_{\zeta_{n-1}}) = \mathbb{P}(\chi_n = i \mid \mathcal{F}_{\chi_{n-1}}).$$

B. Proof of Lemma 2

Lemma 2. (Dominance) *In ALnETC, if the incentive sensitivity function $G(\cdot)$ and the payment b satisfy $G(b, t) > 1$ for all t in the exploration and exploitation phases, then the expected dominant time τ_s is $O(\log T)$.*

Recall that the definition of dominance is at time $t \geq \tau_n$, $S_{\hat{a}^*}(t) \geq \sum_{a \neq \hat{a}^*} S_a(t)$. Thus arm $\hat{a}^*$ is expected to dominate at time $t \geq \tau_n$ if

$$\mu_{\hat{a}^*} \mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a \neq \hat{a}^*} \mu_a \mathbb{E}[T_a(t)].$$

We tighten this condition by narrowing the left-hand-side and amplifying the right-hand-side as follows:

$$\begin{aligned}
&\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t)]\\
&\Rightarrow T_{\hat{a}^*}(\tau_n)+\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)-T_{\hat{a}^*}(\tau_n)] \geq \sum_{a\neq\hat{a}^*}T_a(\tau_n)+\sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t)-T_a(\tau_n)]\\
&\Rightarrow n+\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)-T_{\hat{a}^*}(\tau_n)] \overset{\text{(i)}}{\geq} (\mu_{\hat{a}^*}\mathbb{E}[\tau_n]-n)+\sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t)-T_a(\tau_n)]\\
&\Rightarrow n+\mu_{\hat{a}^*}\frac{G(b,t)}{G(b,t)+1}\mathbb{E}[t-\tau_n] \overset{\text{(ii)}}{\geq} (\mu_{\hat{a}^*}\mathbb{E}[\tau_n]-n)+\mu_{\hat{a}^*}\frac{\mathbb{E}[t-\tau_n]}{G(b,t)+1}\\
&\Rightarrow \mathbb{E}[t-\tau_n] \overset{\text{(iii)}}{\geq} \frac{\big(\mathbb{E}[\tau_n]-\frac{2n}{\mu_{\hat{a}^*}}\big)\big(G(b,t)+1\big)}{G(b,t)-1},
\end{aligned} \tag{3}$$

where (i) is because arm $\hat{a}^*$ is pulled at least n times during the exploration phase, (ii) is because by incentivizing arm $\hat{a}^*$, we have $\hat{\lambda}_{\hat{a}^*}(t) \geq \frac{G(b,t)}{G(b,t)+1}$ and $\hat{\lambda}_a(t) \leq \frac{1}{G(b,t)+1}$ for $a \neq \hat{a}^*$, and (iii) is the rearrangement. Then we obtain the sufficient condition of dominance (3). Since time τ_s is defined as the earliest time to reach dominance, we can upper bound $\mathbb{E}[\tau_s-\tau_n]$ by

$$\mathbb{E}[\tau_s-\tau_n] \leq \frac{\big(\mathbb{E}[\tau_n]-\frac{2n}{\mu_{\hat{a}^*}}\big)\big(G(b,t)+1\big)}{G(b,t)-1}. \tag{4}$$

Next, we prove the following result for $\mathbb{E}[\tau_n]$.

Lemma 6. *In ALnETC, the expected exploration phase duration $\mathbb{E}[\tau_n]$ is upper bounded by $O(\log T)$.*

B.1. Proof of Lemma 6

In ALnETC, during the exploration phase at time step t, the agent offers payment b to the user pulling arm i. The probability that the arm i generates reward is $\frac{\lambda_i(t)+G(b,t)}{1+G(b,t)}\cdot\mu_i > \frac{G(b,t)\mu_i}{1+G(b,t)}$. Thus, the number of attempts for arm i to generate a unit reward is a geometric random variable with parameter larger than $\frac{G(b,t)\mu_i}{1+G(b,t)}$. By the policy, during the exploration phase, each arm generates at least n accumulative reward. Then we obtain

$$\mathbb{E}[\tau_n] \leq n\cdot\sum_{i\in A}\frac{1+G(b,t)}{G(b,t)\mu_i} = O(n) = O(\log T). \tag{5}$$

Lastly, it follows from Lemma 6 that $\mathbb{E}[\tau_s] = \mathbb{E}[\tau_n]+\mathbb{E}[\tau_s-\tau_n] = O(\log T)$. This completes the proof.

C. Proof of Theorem 3

Theorem 3. (At-Least-n Explore-Then-Commit) *Given a fixed time horizon T, if (i) $G(b,t) > 1$, (ii) $q \geq (2\max_{a\neq a^*}\mu_a)/\Delta_{min}^2$, (iii) $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then the expected regret of ALnETC is upper bounded by:*

$$\mathbb{E}[R_T] \leq \sum_{a\in A}\frac{2(G(b,t)-L_{a^*})\Delta_{max}}{\big(G(b,t)-1\big)\mu_a}\cdot q\ln T + o(\log T),$$

where $L_a = F(q\ln T+\theta_a)/\sum_{i\in A}F(\mu^*T+\theta_i)$. The expected total payment is upper bounded by:

$$\mathbb{E}[B_T] \leq \sum_{a\neq a^*}\frac{2b(G(b,t)+1)}{\mu_a(G(b,t)-1)}\cdot q\ln T.$$

In the rest of the proofs, for simplicity we will use the notations $\Delta_a = \mu^*-\mu_a$, $\mu_{min} = \min_{a\in A}\mu_a$, $\Delta_{max} = \max_{a\in A}\Delta_a$ and $\Delta_{min} = \min_{a\in A}\Delta_a$.

By the law of total expectation, the expected regret up to T is as follows:

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}[R_T \mid \hat{a}^* = a^*]\mathbb{P}(\hat{a}^* = a^*) + \mathbb{E}[R_T \mid \hat{a}^* \neq a^*]\mathbb{P}(\hat{a}^* \neq a^*) \\ &\leq \mathbb{E}[R_T \mid \hat{a}^* = a^*] + T \cdot \mathbb{P}(\hat{a}^* \neq a^*).\end{aligned}$$

We want to bound both $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$ and $\mathbb{P}(\hat{a}^* \neq a^*)$ to get the regret bound. First we analyze the upper bound of the part $\mathbb{P}(\hat{a}^* \neq a^*)$. We start with the following lemma.

Lemma 7. *For each arm $a \neq a^*$, there exists a constant $\epsilon_a > 0$ independent of n such that the following hold:*

$$\mathbb{P}\left(\hat{\mu}_a(\tau_n) > \mu_a + \frac{\Delta_a}{2}\right) \leq 2e^{-2\epsilon_a n},$$

and

$$\mathbb{P}\left(\hat{\mu}_{a^*}(\tau_n) < \mu_{a^*} - \frac{\Delta_a}{2}\right) \leq 2e^{-2\epsilon_a n}.$$

Let arm $a = \arg\max_{i \in A, i \neq a^*} \hat{\mu}_i(\tau_n)$ denote the arm with largest sample mean and not equal to arm a^* at time step τ_n. We have:

$$\begin{aligned}\mathbb{P}(\hat{a}^* \neq a^*) &\leq \mathbb{P}\left(\hat{\mu}_a(\tau_n) \geq \hat{\mu}_{a^*}(\tau_n)\right) \\ &\overset{\text{(i)}}{\leq} \mathbb{P}\left(\hat{\mu}_a(\tau_n) \geq \mu_a + \frac{\Delta_a}{2}\right) + \mathbb{P}\left(\hat{\mu}_{a^*}(\tau_n) \leq \mu_{a^*} - \frac{\Delta_a}{2}\right) \\ &\overset{\text{(ii)}}{\leq} 4e^{-\frac{n\Delta_a^2}{2\mu_a}},\end{aligned}$$

where (i) is because $\mu_a + \Delta_a/2 = \mu_{a^*} - \Delta_a/2$, and the event $\{\hat{\mu}_a(\tau_n) \geq \hat{\mu}_{a^*}(\tau_n)\}$ implies either $\{\hat{\mu}_a(\tau_n) \geq \mu_a + \Delta_a/2\}$ or $\{\hat{\mu}_{a^*}(\tau_n) \leq \mu_{a^*} - \Delta_a/2\}$, and (ii) follows by leveraging Lemma 7. Recall that, in the policy, we define $n = q \log T$. Thus, if $q \geq \frac{2 \max_{a \neq a^*} \mu_a}{\Delta_{min}^2}$, it then follows that $\mathbb{P}(\hat{a}^* \neq a^*) = O(\frac{1}{T})$.

Next, we analyze the upper bound of the part $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$. Let Γ_t denote the accumulative reward up to time step t. Then, we have:

$$\begin{aligned}\mathbb{E}[R_T \mid \hat{a}^* = a^*] &= \mathbb{E}[\Gamma_T^*] - \mathbb{E}[\Gamma_T \mid \hat{a}^* = a^*] \\ &= \mu^* \cdot T - \mathbb{E}[\Gamma_T \mid \hat{a}^* = a^*] \\ &= \mu^* \cdot T - \big(\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*] + \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*]\big). \qquad (6)\end{aligned}$$

During the exploration phase, since each arm generates rewards at least n times, we obtain:

$$\begin{aligned}\mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] &= \mathbb{E}\left[\sum_{i \in A}\big(n + (S_i(\tau_n) - n)\big)\right] \\ &= m \cdot n + \mathbb{E}\left[\sum_{i \in A}\big(T_i(\tau_n) \cdot \mu_i - n\big)\right] \\ &= m \cdot n + \sum_{i \in A} \mu_i \left(\mathbb{E}[T_i(\tau_n)] - \frac{n}{\mu_i}\right) \\ &\geq m \cdot n + \mu_{min} \cdot \sum_{i \in A}\left(\mathbb{E}[T_i(\tau_n)] - \frac{n}{\mu_i}\right) \\ &= m \cdot n + \left(\tau_n \cdot \mu_{min} - \mu_{min} \cdot \sum_{i \in A} \frac{n}{\mu_i}\right) \\ &= \tau_n \cdot \mu_{min} + n \cdot \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i}. \qquad (7)\end{aligned}$$

For each arm $a \in A$, let $L_a = \frac{F(q \ln T + \theta_a)}{\sum_{i \in A} F(\mu^* T + \theta_i)}$. Thus at time $t \in \{\tau_n + 1, \ldots, T\}$, we have

$$\mathbb{E}[\lambda_a(t)] = \mathbb{E}\left[\frac{F(S_a(t-1) + \theta_a)}{\sum_{i \in A} F(S_i(t-1) + \theta_i)}\right] \overset{\text{(i)}}{\geq} \frac{F(q \ln T + \theta_a)}{\sum_{i \in A} F(\mu^* T + \theta_i)} = L_a,$$

where (i) is obtained since at time $t > \tau_n$, $S_a(t-1) \geq q \ln T$ and $S_a(t-1) \leq \mu^* T$ for any $a \neq a^*$.

During the exploitation phase, the agent offers payment to users pulling arm $\hat{a}^*$, so using the bound in (7) we obtain:

$$\begin{aligned}
&\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*, \tau_n, \tau_s] \\
&= \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{\lambda_{a^*}(t) + G(b,t)}{1 + G(b,t)} \cdot \mu^* + \sum_{i \in A} \frac{\lambda_i(t)}{1 + G(b,t)} \cdot \mu_i\right] \\
&\geq \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{\lambda_{a^*}(t) + G(b,t)}{1 + G(b,t)} \cdot \mu^* + \frac{(1 - \lambda_{a^*}(t))}{1 + G(b,t)} \cdot \mu_{min}\right] \\
&= \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{G(b,t)}{1 + G(b,t)} \cdot \mu^* + \frac{\mu_{min}}{1 + G(b,t)} + \frac{\lambda_{a^*}(t)\Delta_{max}}{1 + G(b,t)}\right] \\
&\geq \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \frac{\mu^*(\tau_s - \tau_n)G(b,t)}{1 + G(b,t)} + \frac{(\tau_s - \tau_n)\mu_{min}}{1 + G(b,t)} + \frac{(\tau_s - \tau_n)L_{a^*}\Delta_{max}}{1 + G(b,t)} \\
&\overset{\text{(i)}}{\geq} \tau_n \cdot \mu_{min} + n \cdot \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^*(\tau_s - \tau_n)G(b,t)}{1 + G(b,t)} + \frac{(\tau_s - \tau_n)\mu_{min}}{1 + G(b,t)} + \frac{(\tau_s - \tau_n)L_{a^*}\Delta_{max}}{1 + G(b,t)} \\
&= n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1 + G(b,t)}\tau_s + \tau_n \cdot \mu_{min} - \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1 + G(b,t)}\tau_n \\
&= n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1 + G(b,t)}\tau_s - \frac{\big(G(b,t) + L_{a^*}\big)\Delta_{max}}{1 + G(b,t)}\tau_n,
\end{aligned} \tag{8}$$

where (i) is obtained by replacing $\mathbb{E}[\Gamma_{\tau_n} \mid \tau_n]$ using (7). Then replacing (6) using (8) and taking expectation with respect to τ_n and τ_s, we obtain:

$$\begin{aligned}
&\mathbb{E}[R_T \mid \hat{a}^* = a^*] \\
&\leq \mu^* T - \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1 + G(b,t)}\mathbb{E}[\tau_s] + \frac{\big(G(b,t) + L_{a^*}\big)\Delta_{max}}{1 + G(b,t)}\mathbb{E}[\tau_n] - n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} \\
&\quad - \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*] \\
&= \mu^* \mathbb{E}[\tau_s] - \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1 + G(b,t)}\mathbb{E}[\tau_s] + \frac{\big(G(b,t) + L_{a^*}\big)\Delta_{max}}{1 + G(b,t)}\mathbb{E}[\tau_n] - n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} \\
&\quad + \mu^*\big(T - \mathbb{E}[\tau_s]\big) - \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*] \\
&= \frac{\Delta_{max}(1 - L_{a^*})}{1 + G(b,t)}\mathbb{E}[\tau_s - \tau_n] + \Delta_{max} \cdot \mathbb{E}[\tau_n] - n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*].
\end{aligned} \tag{9}$$

Then, the evaluation of $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$ boils down to evaluating $\mathbb{E}[\tau_n]$, $\mathbb{E}[\tau_s - \tau_n]$ and $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$. We obtain

from Lemma 2 and (9) that

$$
\begin{aligned}
&\mathbb{E}[R_T \mid \hat{a}^* = a^*] \\
&\leq \frac{\Delta_{max}(1-L_{a^*})}{1+G(b,t)} \cdot \frac{\big(n \cdot \sum_{i\in A} \frac{1+G(b,t)}{G(b,t)\mu_i} - \frac{2n}{\mu^*}\big)\big(G(b,t)+1\big)}{G(b,t)-1} + \Delta_{max} n \sum_{i\in A} \frac{1+G(b,t)}{G(b,t)\mu_i} - n\sum_{i\in A}\frac{\mu_i-\mu_{min}}{\mu_i} \\
&\quad + \mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*] \\
&= n\left[\frac{\big(G(b,t)-L_{a^*}\big)\big(G(b,t)+1\big)}{G(b,t)\big(G(b,t)-1\big)}\sum_{a\in A}\frac{\Delta_{max}}{\mu_a} - \frac{a\Delta_{max}(1-L_{a^*})}{\mu^*\big(G(b,t)-1\big)} - \sum_{a\in A}\frac{\mu_a-\mu_{min}}{\mu_a}\right] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*] \\
&\overset{\text{(i)}}{\leq} n\left[\frac{2\big(G(b,t)-L_{a^*}\big)}{G(b,t)-1}\sum_{a\in A}\frac{\Delta_{max}}{\mu_a}\right] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*] \\
&= O(\log T) + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*],
\end{aligned}
$$

where (i) follows because $G(b,t)+1 < 2G(b,t)$. By leveraging Eqs (5) and (4), the expected accumulative payment $\mathbb{E}[B_T]$ can also be upper bounded by

$$
\mathbb{E}[B_T] = b \cdot (\mathbb{E}[\tau_n] + \mathbb{E}[\tau_s - \tau_n]) \leq \sum_{a\neq a^*} \frac{2b(G(b,t)+1)}{\mu_a(G(b,t)-1)} \cdot q \ln T = O(\log T).
$$

Next, for simplicity, we consider a system with $A = \{1,2\}$, where $\mu_1 > \mu_2$ and $\theta_1, \theta_2 > 0$. The idea of the policy is that the agent keeps offering payment b to the users pulling arm 1 to help accumulate reward from arm 1 and keep the arm in the leading side, i.e., arm 1 generates at least half of accumulative reward, until time step τ_s when arm 1 dominates and has an overwhelming chance to be the only arm that can generate rewards after monopoly happens. This phenomenon is formulated as follows: suppose at time step τ_s, $S_1(\tau_s) + S_2(\tau_s) = n_0$, and $S_2(\tau_s) = u_0 n_0$ with $0 < u_0 < \frac{1}{2}$ and $u_0 n_0 \gg \theta_1, \theta_2$. We estimate the probability of a "bad" event $D(u_0, n_0)$, where at some time step $t' > \tau_s$ we have $S_1(t') + S_2(t') = n > n_0$ and $S_2(t') \geq un$ with $0 < u_0 < u < \frac{1}{2}$, by leveraging the improved exponential embedding method, $D(u_0, n_0)$ can be expressed as follows:

$$
D(u_0, n_0) = \left(\sum_{i=u_0 n_0}^{un-1} r_2(i) < \sum_{i=n_0 - u_0 n_0}^{n-un-1} r_1(i) \right).
$$

We will show later that $\mathbb{P}(D(u_0, n_0))$ is very small, and with $u_0 n_0$ getting larger, $\mathbb{P}(D(u_0, n_0))$ is getting exponentially smaller. This result is formally stated as follows:

Lemma 8. *Suppose at time step τ_s there are n_0 accumulative reward with $u_0 n_0, 0 < u_0 < \frac{1}{2}$ generated by arm 2. Then, there exists a constant $\gamma \in (0, 1/4)$, such that for any $u_0 < u < \frac{1}{2}$ and all large enough n_0, it holds that:*

$$
\mathbb{P}\Big(\exists n > n_0, D(u_0, n_0)\Big) \leq e^{-(u_0 n_0)^\gamma}.
$$

By the above lemma, with $u_0 n_0 = O(\tau_n) = O(\log T)$, we get $\mathbb{P}\big(D(u_0, n_0)\big) = O(e^{-(\log T)^\gamma})$. This result can be extended to the case with arm number $m \geq 2$, by viewing the sum of accumulative reward generated from all sub-optimal arms as the accumulative reward generated from a single "super arm."

Next, we bound the last part $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$. Note that the regret comes from pullings of sub-optimal arms, and the expected number of attempts for each arm to get a unit reward is $O(1)$ since $\mu_i > 0, i \in A$. Let n_0 denote the accumulative reward from all arms at time step τ_s with $u_0 n_0, 0 < u_0 < \frac{1}{2}$ rewards generated by sub-optimal arms. Note that $u_0 n_0 = O(\log T)$ since $u_0 n_0 < \tau_s$ and $\tau_s = O(\log T)$. Then, by Lemma 8, for the unit reward generated right after τ_s, it is generated by sub-optimal arms with probability smaller than or equal to $e^{-(u_0 n_0)^\gamma}$ with $\gamma \in (0, \frac{1}{4})$. When a unit reward is generated by sub-optimal arms, the probability that the next unit reward is also generated by sub-optimal arms is smaller

than or equal to $e^{-(u_0n_0+1)^\gamma}$. Thus, we can upper bound the expected regret $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$ by

$$\begin{aligned}\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*] &\le e^{-(u_0n_0)^\gamma} + e^{-(u_0n_0+1)^\gamma} + \cdots \\ &\le \int_{u_0n_0-1}^{\infty} e^{-n^\gamma} dn \\ &= Ce^{-(u_0n_0-1)^\gamma}, \end{aligned} \tag{10}$$

where C only depends on u_0n_0 and γ such that $C = O\big((u_0n_0)^{1-\gamma}\big)$ with $\gamma \in (0, 1/4)$. Thus Eq. (10) is $o(\log T)$. Now we get the expected regret up to time step T as $\mathbb{E}[R_T] = O(\log T)$, this completes the proof.

C.1. Proof of Lemma 7

Fact 3 (Chernoff-Hoeffding bound)**.** *Let $Z_1, \cdots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, where $-\infty < a \le b < \infty$. Then for all $s \ge 0$*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| \ge s\right) \le \exp\left(-\frac{2ns^2}{(b-a)^2}\right).$$

Let sequences $\{X_i(t)\}$ denote the Bernoulli reward with support $\{0, 1\}$ generated by arm $i \ne a^*$ at time step t. Thus, for each time step t, $X_i(t)$ is an i.i.d. random variable and $\mathbb{E}[X_i(t)] = \mu_i$. At time step τ_n, by the policy, each arm has at least n accumulative reward. Since $S_i(\tau_n)$ is the accumulative reward generated by arm i at time step τ_n we have $S_i(\tau_n) \ge n$. By Chernoff-Hoeffding bound, at time step τ_n for arm i, we get the following:

$$\mathbb{P}\left(\hat{\mu}_i(\tau_n) > \mu_i + \frac{\Delta_i}{2}\right) \le 2e^{-2\mathbb{E}[T_i(\tau_n)](\frac{\Delta_i}{2})^2} = 2e^{-2\frac{\mathbb{E}[S_i(\tau_n)]}{\mu_i}(\frac{\Delta_i}{2})^2} \le 2e^{-\frac{n\Delta_i^2}{2\mu_i}}.$$

The proof for arm a^* also follows from similar arguments and thus is omitted for brevity.

C.2. Proof of Lemma 8

Suppose at some time step t, there are n accumulative reward from both arms. Recall that for arm $i \in A$, $\sum_{j=n}^{\infty} r_i(j) < \infty$ and $\mathbb{E}\big[\sum_{j=n}^{\infty} r_i(j)\big] = \sum_{j=n}^{\infty} \frac{1}{\mu_i F(j+\theta_i)}$ converges. To prove Lemma 8, we use the following lemma

Lemma 9. *There exists a constant n_0 such that for all $n > n_0$,*

$$\mathbb{P}\left(\left|\frac{\sum_{j=n}^{\infty} r_i(j)}{\mathbb{E}\big[\sum_{j=n}^{\infty} r_i(j)\big]} - 1\right| > n^{-\frac{1}{4}}\right) \le e^{-n^{\frac{1}{4}}}, i \in A.$$

Given a constant t, define an event E_{n_0} where the following conditions hold simultaneously:

$$\left|\frac{\sum_{j=u_0n_0}^{\infty} r_2(j)}{\mathbb{E}\big[\sum_{j=u_0n_0}^{\infty} r_2(j)\big]} - 1\right| \le (u_0n_0)^{-\frac{1}{4}}, \tag{11}$$

$$\forall n > n_0, \left|\frac{\sum_{j=un}^{\infty} r_2(j)}{\mathbb{E}\big[\sum_{j=un}^{\infty} r_2(j)\big]} - 1\right| \le (un)^{-\frac{1}{4}}, \tag{12}$$

$$\left|\frac{\sum_{j=(1-u_0)n_0}^{\infty} r_1(j)}{\mathbb{E}\big[\sum_{j=(1-u_0)n_0}^{\infty} r_1(j)\big]} - 1\right| \le \big((1-u_0)n_0\big)^{-\frac{1}{4}}, \tag{13}$$

$$\forall n > n_0, \left|\frac{\sum_{j=(1-u)n}^{\infty} r_1(j)}{\mathbb{E}\big[\sum_{j=(1-u)n}^{\infty} r_1(j)\big]} - 1\right| \le \big((1-u)n\big)^{-\frac{1}{4}}. \tag{14}$$

By Lemma 9, we obtain the probability of event E_{n_0} as follows

$$\mathbb{P}(E_{n_0}) \ge 1 - 2e^{-(u_0n_0)^{\frac{1}{4}}} - \sum_{n>n_0} 2e^{-(u_0n)^{\frac{1}{4}}} \ge 1 - e^{-(u_0n_0)^\gamma},$$

with $\gamma \in (0, \frac{1}{4})$ depending only on F and u_0. If we show that for all large enough $u_0 n_0$, $E_{n_0} \cap D(u_0, n_0) = 0$, then the proof is finished since it implies

$$\mathbb{P}\bigg(\exists n > n_0, D(u_0, n_0)\bigg) \leq \mathbb{P}(E_{n_0}^c) \leq e^{-(u_0 n_0)^\gamma}.$$

We consider the definition of event $D(u_0, n_0)$. By (11)–(14), we obtain

$$\begin{aligned}\sum_{i=u_0 n_0}^{un-1} r_2(i) &= \sum_{i=u_0 n_0}^{\infty} r_2(i) - \sum_{i=un}^{\infty} r_2(i) \\ &\geq \big(1+o(1)\big) \sum_{i=u_0 n_0}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} - \big(1+o(1)\big) \sum_{i=un}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)},\end{aligned}$$

and similarly,

$$\sum_{i=n_0-u_0 n_0}^{n-un-1} r_1(i) \leq \big(1+o(1)\big) \sum_{i=(1-u_0)n_0}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)} - \big(1+o(1)\big) \sum_{i=(1-u)n}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)}.$$

By contradiction, suppose that $E_{n_0} \cap D(u_0, n_0) \neq 0$. It then follows that

$$\begin{aligned}&\big(1+o(1)\big) \sum_{i=u_0 n_0}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} - \big(1+o(1)\big) \sum_{i=un}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} \\ &< \big(1+o(1)\big) \sum_{i=(1-u_0)n_0}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)} - \big(1+o(1)\big) \sum_{i=(1-u)n}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)},\end{aligned}$$

which implies

$$\sum_{i=u_0 n_0}^{(1-u_0)n_0} \frac{1}{\mu_1 F(i+\theta_1)} < \big(1+o(1)\big) \sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(i+\theta_1)}. \tag{15}$$

We want to show that (15) cannot hold as $u_0 n_0$ goes large, which implies $E_{n_0} \cap D(u_0, n_0) = 0$. Since $F(x) = \Omega(x^\alpha)$, there exists $k > 0$ such that

$$\begin{aligned}\sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(i+\theta_1)} &\leq k\bigg(\frac{n_0}{n}\bigg)^\alpha \sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(\frac{n_0}{n} i + \frac{n_o}{n}\theta_1)} \\ &= k\bigg(\frac{n_0}{n}\bigg)^\alpha \sum_{i=un_0}^{(1-u)n_0} \frac{1}{\mu_1 F(i+\theta_1)}.\end{aligned}$$

Also, note that $[un_0, (1-u)n_0] \subset [u_0 n_0, (1-u_0)n_0]$. Therefore, there exists a constant $d \in (0,1)$ such that

$$\sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(i+\theta_1)} \leq dk\bigg(\frac{n_0}{n}\bigg)^\alpha \sum_{i=u_0 n_0}^{(1-u_0)n_0} \frac{1}{\mu_1 F(i+\theta_1)},$$

which contradicts with (15) since $o(1)$ goes to 0 as $u_0 n_0$ goes to infinity, and this completes the proof.

C.3. Proof of Lemma 9

Let $R_n = \sum_{j=n}^{\infty} r_i(j)$, $h(j) = \mu_i F(j+\theta_i)$, $Z_n = \sum_{j=n}^{\infty} \frac{1}{h(j)^2}$. We first show that for any $t \in \mathbb{R}^+$, we have

$$\mathbb{P}(R_n - \mathbb{E}[R_n] > t\sqrt{Z_n}) \leq e^{-t}, \tag{16}$$

and

$$\mathbb{P}(R_n - \mathbb{E}[R_n] < -t\sqrt{Z_n}) \leq e^{-t}. \tag{17}$$

We only prove the first inequality and the proof of the second one is similar. Given a constant s, we have:

$$\begin{aligned}
\mathbb{P}(R_n - \mathbb{E}[R_n] > t\sqrt{Z_n}) &\overset{\text{(i)}}{=} \mathbb{P}\left(e^{s(R_n - \mathbb{E}[R_n])} > e^{st\sqrt{Z_n}}\right) \\
&\overset{\text{(ii)}}{\leq} e^{-st\sqrt{Z_n}}\mathbb{E}\left[e^{s\sum_{j\geq n}(r_i(j) - \frac{1}{h(j)})}\right] \\
&= e^{-st\sqrt{Z_n}}\prod_{j\geq n}\mathbb{E}\left[e^{s(r_i(j) - \frac{1}{h(j)})}\right] \\
&\overset{\text{(iii)}}{=} e^{-st\sqrt{Z_n}}\prod_{j\geq n}\frac{e^{-\frac{s}{h(j)}}}{1 - \frac{s}{h(j)}} \\
&= e^{-st\sqrt{Z_n}}\prod_{j\geq n} e^{\frac{-s}{h(j)}}\left[1 + \frac{s}{h(j)} + \frac{\frac{s^2}{h(j)^2}}{1 - \frac{s}{h(j)}}\right] \\
&\overset{\text{(iv)}}{\leq} e^{-st\sqrt{Z_n}}\prod_{j\geq n} e^{\frac{2s^2}{h(j)^2}} \\
&\leq \exp(2s^2 Z_n - st\sqrt{Z_n}),
\end{aligned} \tag{18}$$

where (i) follows from multiplying both sides by a variable s and exponentiate both sides, (ii) follows from Markov's inequality, (iii) is because given random variable $X \sim Exp(\lambda)$, $\mathbb{E}[e^{aX}] = \frac{1}{1-\frac{a}{\lambda}}, a < \lambda$, and (iv) follows from $e^x \geq 1 + x$. We set $s = \frac{1}{\sqrt{Z_n}}$, which is achievable since there exists n such that $\frac{1}{\sqrt{Z_n}} \leq \frac{h(n)}{2}$. Thus, by (18), we obtain $\mathbb{P}(R_n - \mathbb{E}[R_n] > t\sqrt{Z_n}) \leq e^{-t}$. Next, we use Lemma 1 in Oliveira (2009), which is restated as follows:

Lemma 10 (Oliveira (2009), Lemma 1)**.** *Define a feedback function $F(x) = \Theta(x^\alpha)$ where $\alpha > 1$, and define the quantity*

$$S_r(n) = \sum_{j=n}^{\infty}\frac{1}{F(j)^r}, r \in \mathbb{R}^+, n \in \mathbb{N}.$$

Then, for all $r \geq 1$, $S_r(n)$ converges and as $n \to +\infty$

$$S_r(n) \to \frac{n}{(r\alpha - 1)F(n)^r}.$$

By using Lemma 10, we obtain $\sqrt{S_2(n)} = n^{-\frac{1}{2}}S_1(n)$ asymptotically. Note that $S_1(n) = \mu_i\mathbb{E}[R_n]$ and $S_2(n) = \mu_i^2 Z_n$. Therefore, we obtain the relation between $\mathbb{E}[R_n]$ and $\sqrt{Z_n}$ as $\sqrt{Z_n} = n^{-\frac{1}{2}}\mathbb{E}R_n$ asymptotically. Then we replace t by $n^{\frac{1}{4}}$ in both (16) and (17), and we get the inequality in Lemma 9.

D. Proof of Theorem 4

Theorem 4. (UCB-List) *Given a fixed time horizon T, if $G(b,t) > 1$, and $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then the expected regret of UCB-List $\mathbb{E}[R_T]$ is upper bounded by*

$$\sum_{a\neq a^*}\left[\frac{8\Delta_a\big(G(b,t)-1\big)+8\Delta_{max}}{\big(G(b,t)-1\big)\Delta_a^2}\ln T + 4\Delta_a + \frac{4\Delta_{max}}{G(b,t)-1}\right],$$

with the expected payment $\mathbb{E}[B_T]$ upper bounded by

$$\frac{2G(b,t)+1}{G(b,t)-1}\left[\frac{8b\ln T}{\Delta_{min}^2} + \sum_{a\neq a^*}\left(\frac{8b\ln T}{\Delta_a^2} + 4b\right)\right].$$

We start in a similar way as the proof of Theorem 3. By the law of total expectation, the expected regret up to T can be bounded as follows:

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}[R_T \mid \hat{a}^* = a^*]\mathbb{P}(\hat{a}^* = a^*) + \mathbb{E}[R_T \mid \hat{a}^* \neq a^*]\mathbb{P}(\hat{a}^* \neq a^*) \\ &\leq \mathbb{E}[R_T \mid \hat{a}^* = a^*] + T \cdot \mathbb{P}(\hat{a}^* \neq a^*).\end{aligned}$$

We want to bound both $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$ and $\mathbb{P}(\hat{a}^* \neq a^*)$ to get the regret bound. We first consider $\mathbb{E}[R_T \mid \hat{a}^* = a^*]$. After decomposing, we have:

$$\begin{aligned}\mathbb{E}[R_T \mid \hat{a}^* = a^*] &= \mathbb{E}[R_{\tau_2} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*] \\ &= \mathbb{E}[R_{\tau_1}] + \mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*]. \end{aligned} \tag{19}$$

Note that after initialization, i.e., let t_0 be the time step when initialization is finished, each arm a has $T_a(t_0) \geq 1$ since the number of attempts for each arm a to get a unit reward is a geometric random variable with parameter larger than $\frac{G(b,t)\mu_a}{1+G(b,t)}$, which is independent of time. During the exploration phase, since the regret is caused by pullings of sup-optimal arms, the expected regret after t time steps can be written as

$$\sum_{a \neq a^*, a \in A} \Delta_a \mathbb{E}[T_a(t)].$$

Thus we can bound the expected regret during the exploration phase $\mathbb{E}[R_{\tau_1}]$ by bounding each $\mathbb{E}[T_a(\tau_1)]$ for $a \neq a^*$. Let $U(t)$ denote the set of arms that can get payment at time t. Consider the following two cases during the exploration phase:

(a) At time $t \leq \tau_1$, $a^* \in U(t)$ and there exists at least one suboptimal arm $a \in A, a \neq a^*$ such that $a \in U(t)$. Recall that $c_a(t) = \sqrt{\ln T / 2T_a(t)}$ is the confidence bound of arm a at time step. In this case, we have:

$$\begin{aligned}\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \in U(t)\big) &\overset{\text{(i)}}{\leq} \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big) \cdot \mathbb{P}\big(\hat{\mu}^*(t) + c_{a^*}(t) > \hat{\mu}_a(t) - c_a(t)\big) \\ &\leq \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big) \\ &\overset{\text{(ii)}}{\leq} \mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) + \mathbb{P}\left(\hat{\mu}^*(t) - c_{a^*}(t) < \mu^* - \frac{\Delta_a}{2}\right), \end{aligned} \tag{20}$$

where (i) is obtained since arm $a, a^* \in U(t)$ implies that the upper confidence bound of both arms is larger than the other arms's lower confidence bound, (ii) is because $\mu_a + \Delta_a/2 = \mu^* - \Delta_a/2$, and the event $\{\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\}$ implies either $\{\hat{\mu}_a(t) + c_a(t) > \mu_a + \Delta_a/2\}$ or $\{\hat{\mu}^*(t) < \mu^* - \Delta_a/2\}$. We consider the first probability in Eq. (20). By Chernoff-Hoeffding bound we have

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) &= \mathbb{P}\left(\hat{\mu}_a(t) - \mu_a > \frac{\Delta_a}{2} - c_a(t)\right) \\ &\leq e^{-2T_a(t)\left(\frac{\Delta_a}{2} - c_a(t)\right)^2} \\ &= e^{-\left(\ln T + \frac{\Delta_a^2}{2} T_a(t) - \Delta_a \sqrt{2T_a(t) \ln T}\right)}. \end{aligned} \tag{21}$$

Let $\frac{\Delta_a^2}{2} T_a(t) - \Delta_a \sqrt{2T_a(t) \ln T} = 0$, we obtain $T_a(t) = 8 \ln T / \Delta_a^2$ and Eq. (21) equals $1/T$. Note that as $T_a(t)$ increases, Eq. (21) decreases monotonically. Similar bound can be obtained of the second probability in Eq. (20). Thus, in this case, the expected regret contributed by a suboptimal arm $a \in A$ is bounded by

$$\begin{aligned}\Delta_a \mathbb{E}[T_a(t)] &\leq \frac{8 \ln T}{\Delta_a} + \Delta_a T \cdot \mathbb{P}\big(t < \tau_1 : a \in U(t), a^* \in U(t)\big) \\ &\leq \frac{8 \ln T}{\Delta_a} + 2\Delta_a. \end{aligned} \tag{22}$$

(b) At time $t \leq \tau_1$, a^* is eliminated by some suboptimal arm $a \in U(t), a \neq a^*$. In this case, with similar technique as that

in case **(a)** and Chernoff-Hoeffding bound, we have

$$\begin{aligned}
\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \notin U(t)\big) &\leq \mathbb{P}\big(\hat{\mu}_a(t) - c_a(t) > \hat{\mu}^*(t) + c_{a^*}(t)\big) \\
&\leq \mathbb{P}\big(\hat{\mu}_{a^*}(t) + c_{a^*}(t) \leq \mu_{a^*} - \frac{\Delta_a}{2}\big) + \mathbb{P}\big(\hat{\mu}_a(t) - c_a(t) \geq \mu_a + \frac{\Delta_a}{2}\big) \\
&\leq e^{-2T_{a^*}(t)\left(\frac{\Delta_a}{2} + c_{a^*}(t)\right)^2} + e^{-2T_a(t)\left(\frac{\Delta_a}{2} + c_a(t)\right)^2} \\
&= e^{-\frac{\Delta_a^2}{2}T_{a^*}(t) - \ln T - \Delta_a\sqrt{2T_{a^*}(t)\ln T}} + e^{-\frac{\Delta_a^2}{2}T_a(t) - \ln T - \Delta_a\sqrt{2T_a(t)\ln T}} \\
&\leq 2T^{-1}.
\end{aligned}$$

Note that $\mathbb{P}(\hat{a}^* \neq a^*) = \mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \notin U(t)\big)$. Thus, in this case the expected regret contributed by a suboptimal arm $a \in A$ is upper bounded by

$$\Delta_a \mathbb{E}[T_a(t)] \leq \Delta_a T \cdot \mathbb{P}\big(a \in U(t), a^* \notin U(t)\big) = 2\Delta_a. \tag{23}$$

Summing Eq. (22) and Eq. (23) over all suboptimal arms, the expected regret during the exploration phase is bounded by:

$$\mathbb{E}[R_{\tau_1}] \leq \sum_{a \neq a^*} \frac{8\ln T}{\Delta_a} + 4\Delta_a.$$

During the exploration phase at time step $t < \tau_1$, since the agent offers payment b to the user for pulling arm i, the probability that the arm i is pulled is $\frac{\lambda_i(t)+G(b,t)}{1+G(b,t)} > \frac{G(b,t)}{1+G(b,t)}$. Thus, the number of attempts for arm i to get pulled is a geometric random variable with parameter at least $\frac{G(b,t)}{1+G(b,t)}$. Since the above cases **(a)** and **(b)** imply the requirement of $\frac{8\ln T}{\Delta_a^2} + 4$ expected number of pullings from suboptimal arms, thus, the expected number of pullings for a suboptimal arm a to guarantee at most $\frac{8\ln T}{\Delta_a^2} + 4$ number of pullings on every suboptimal arm is upper bounded by:

$$\mathbb{E}[T_a(\tau_1)] \leq \frac{G(b,t)+1}{G(b,t)}\left(\frac{8\ln T}{\Delta_a^2} + 4\right).$$

Thus, $\mathbb{E}[\tau_1]$ is upper bounded by:

$$\mathbb{E}[\tau_1] = \sum_{a \in A} \mathbb{E}[T_a(\tau_1)] \overset{\text{(i)}}{\leq} \frac{G(b,t)+1}{G(b,t)}\left(\frac{8\ln T}{\Delta_{min}^2} + \sum_{a \neq a^*}\big(\frac{8\ln T}{\Delta_a^2} + 4\big)\right), \tag{24}$$

where (i) is due to the requirement of $T_{a^*}(\tau_1)$ to be at most $\frac{8\ln T}{\Delta_{min}^2}$, since the exploration phase stops once the sampled strongest suboptimal arm is eliminated. By the definition of dominance, arm $\hat{a}^*$ is expected to dominate at time $t \geq \tau_1$ if

$$\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a \neq \hat{a}^*} \mu_a \mathbb{E}[T_a(t)].$$

Similar as that in the proof of Lemma 2, after tightening the condition by narrowing the left-hand-side and amplifying the right-hand-side, we obtain the sufficient condition of dominance as follows:

$$\begin{aligned}
&\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a \neq \hat{a}^*} \mu_a \mathbb{E}[T_a(t)] \\
&\Rightarrow \mu_{\hat{a}^*} T_{\hat{a}^*}(\tau_1) + \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - T_{\hat{a}^*}(\tau_1)] \geq \sum_{a \neq \hat{a}^*} \mu_a T_a(\tau_1) + \sum_{a \neq \hat{a}^*} \mu_a \mathbb{E}[T_a(t) - T_a(\tau_1)] \\
&\Rightarrow \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - T_{\hat{a}^*}(\tau_1)] \overset{\text{(i)}}{\geq} \sum_{a \neq \hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big) + \sum_{a \neq \hat{a}^*} \mu_a \mathbb{E}[T_a(t) - T_a(\tau_1)] \\
&\Rightarrow \frac{\mu_{\hat{a}^*} G(b,t)\mathbb{E}[t - \tau_1]}{G(b,t)+1} \overset{\text{(ii)}}{\geq} \sum_{a \neq \hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big) + \frac{\max_{a \neq \hat{a}^*} \mu_a \mathbb{E}[t - \tau_1]}{G(b,t)+1} \\
&\Rightarrow \mathbb{E}[t - \tau_1] \overset{\text{(iii)}}{\geq} \frac{G(b,t)+1}{\mu_{\hat{a}^*} G(b,t) - \max\limits_{a \neq \hat{a}^*}\mu_a} \sum_{a \neq \hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big),
\end{aligned} \tag{25}$$

where (i) is obtained since $T_{\hat{a}^*}(\tau_1) > 0$, (ii) is because by incentivizing arm $\hat{a}^*$, we have $\hat{\lambda}_{\hat{a}^*}(t) \geq \frac{G(b,t)}{G(b,t)+1}$ and $\hat{\lambda}_a(t) \leq \frac{1}{G(b,t)+1}$ for $a \neq \hat{a}^*$, and (iii) is the rearrangement. Since time τ_2 is defined as the earliest time to reach dominance, we can upper bound $\mathbb{E}[\tau_2 - \tau_1]$ by

$$\mathbb{E}[\tau_2 - \tau_1] \leq \frac{G(b,t)+1}{\mu_{\hat{a}^*} G(b,t) - \max_{a \neq \hat{a}^*} \mu_a} \sum_{a \neq \hat{a}^*} \big(\frac{8\mu_a}{\Delta_a^2} \ln T + 4\mu_a\big). \tag{26}$$

Thus, we can bound the regret during the exploitation phase $\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]$ in (19) by

$$\begin{aligned}
\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*] &\overset{\text{(i)}}{\leq} \frac{\Delta_{max}}{G(b,t)+1} \cdot \mathbb{E}[\tau_2 - \tau_1] \\
&\leq \sum_{a \neq a^*} \left(\frac{8\Delta_{max}}{\Delta_a^2 (G(b,t)-1)} \log T + \frac{4\Delta_{max}}{G(b,t)-1} \right),
\end{aligned}$$

where (i) follows because during the exploitation phase there is always a positive probability $\hat{\lambda}_a(t)$ which is at most $\frac{1}{G(b,t)+1}$ to pull suboptimal arm a. By using Eqs (24) and (26), the expected accumulative payment $\mathbb{E}[B_T]$ can also be upper bounded by

$$\begin{aligned}
\mathbb{E}[B_T] &= (\mathbb{E}[\tau_1] + \mathbb{E}[\tau_s - \tau_1]) \cdot b \\
&\leq \frac{G(b,t)+1}{G(b,t)} \left(\frac{8b \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \big(\frac{8b \ln T}{\Delta_a^2} + 4b\big) \right) + \frac{G(b,t)+1}{\mu_{\hat{a}^*} G(b,t) - \max_{a \neq \hat{a}^*} \mu_a} \sum_{a \neq \hat{a}^*} \left(\frac{8b\mu_a}{\Delta_a^2} \ln T + 4b\mu_a \right) \\
&\overset{\text{(i)}}{\leq} \frac{G(b,t)+1}{G(b,t)} \left(\frac{8b \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \big(\frac{8b \ln T}{\Delta_a^2} + 4b\big) \right) + \frac{G(b,t)+1}{G(b,t)-1} \sum_{a \neq \hat{a}^*} \left(\frac{8b}{\Delta_a^2} \ln T + 4b \right) \\
&= \frac{G(b,t)+1}{G(b,t)} \cdot \frac{8b \ln T}{\Delta_{min}^2} + \left(\frac{G(b,t)+1}{G(b,t)} + \frac{G(b,t)+1}{G(b,t)-1} \right) \cdot \sum_{a \neq \hat{a}^*} \left(\frac{8b}{\Delta_a^2} \ln T + 4b \right) \\
&\overset{\text{(ii)}}{\leq} \frac{2G(b,t)+1}{G(b,t)-1} \left[\frac{8b \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \left(\frac{8b \log T}{\Delta_a^2} + 4b \right) \right],
\end{aligned}$$

where (i) follows from $\mu^* > \mu_a$ for $a \neq a^*$, and (ii) follows from rearranging of the coefficients containing $G(b,t)$. The choice of τ_2 is sufficient to make the sampled best arm dominate at time step τ_2 and have overwhelming probability to stay in leading side in monopoly after τ_2. The proof is the same as that in the proof of Theorem 3. Thus, the expected regret of the last part $\mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*] = O((\log T)^{1-\gamma} e^{-(\log T)^\gamma}) = o(\log T)$ with $\gamma \in (0, \frac{1}{4})$ and the proof is the same as that in the proof of Theorem 3.

The above results show that we get the expected regret up to time step T as $\mathbb{E}[R_T] = O(\log T)$ with expected accumulative payment $\mathbb{E}[B_T] = O(\log T)$, which completes the proof.