Bandit Learning Problems in Recommendation Systems: Self-Reinforcing User Preferences, Delayed Feedback, and Online Learning to Rank

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy
in the Graduate School of The Ohio State University

By
Tianchen Zhou, M.S.
Graduate Program in Electrical and Computer Engineering

The Ohio State University
2023

Dissertation Committee:
Jia Liu, Adviser
Yingbin Liang
Atilla Eryilmaz
Xueru Zhang

Abstract

Recommendation systems are playing an increasingly important role in human society, which can be theoretically regarded as a sequential decision making problem and formulated by multi-armed bandit framework. In this thesis, we formulate commonly seen user behavior in recommendation systems, and propose efficient policies in maximizing cumulative reward. First, we investigate a new online learning model that considers real-world phenomena in many recommendation 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. Then, we additionally consider the impact of delayed feedback on previous model. A major challenge of this updated MAB framework is that the loss of information caused by feedback delay complicates both user preference evolution and arm incentivizing decisions, both of which are already highly non-trivial even by themselves. In our analysis, we consider delayed feedbacks that can have either arm-independent or arm-dependent distributions. In both cases, we allow unbounded support for the random delays, i.e., the random delay can be infinite. Next, we study multi-armed bandit

based online learning to rank problem, where the user feedback is partial observable and is generated from two unknown parameters: position preferences and arm means, which has a wide applications such as search engines, video streaming services, and recommender systems in e-commerce. The proposed model considers the setting of multiple user types with different action patterns, and the ranking policies are designed in two strategies: personalized ranking for user experience maximization and equal treatment ranking for fairness, both of which are widely applied in practice. For all the proposed model, we propose and analyze theoretically efficient policies, whose performances are verified by synthetic and real-world experiments.

Vita

2016 ..... Bachelor of Science, Computer Science
Wuhan University
2018 ..... Master of Science, Computer Science
Iowa State University

Publications

Zhou, T., Liu, J., Dong, C. and Deng, J., 2021, July. Incentivized bandit learning with self-reinforcing user preferences. In International Conference on Machine Learning (pp. 12824-12834). PMLR.

Zhou, T., Liu, J., Dong, C. and Sun, Y., 2022. Bandit Learning with Joint Effect of Incentivized Sampling, Delayed Sampling Feedback, and Self-Reinforcing User Preferences. Proc. ICLR.

Zhou, T., Momma, M., Dong, C., Yang, F., Guo, C., Shang, J. and Liu, J.K., 2023. Multi-task learning on heterogeneous graph neural network for substitute recommendation.

Zhou, T., Liu, J., Jiao, Y., Dong, C., Chen, Y., Gao, Y. and Sun, Y., 2023. Bandit Learning to Rank with Position-Based Click Models: Personalized and Equal Treatments. arXiv preprint arXiv:2311.04528.

Fields of Study

Electrical and Computer Engineering

Online statistical learning; Multi-armed bandits; Reinforcement learning.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A multi-armed bandit problem (or bandit problem) is a sequential decision making problem with an exploration-exploitation trade-off. The trade-off is the balance between staying exploiting the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future. In recent years, the bandit framework has received a significant amount of interest in the learning research community. This is partly due to the fact that, in many online e-commerce recommendation systems (e.g., Amazon and Walmart), the problem of online learning of the optimal products while making profits at the same time can be well formulated by a bandit problem. Typically, a three-party bandit framework is usually leveraged to formulate a recommendation system, including platform as the learner, users, and the items (in some situations the items can be third-party vendors), as described in Figure 1.1.

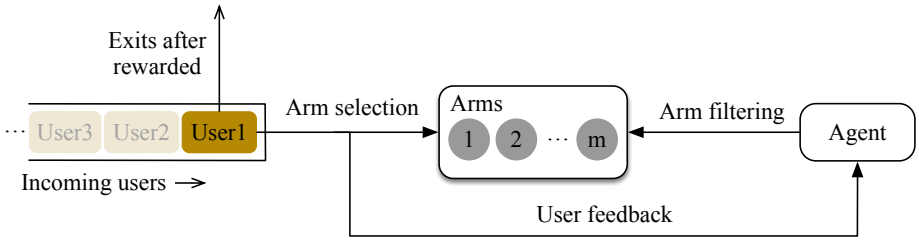


Figure 1.1: A three-party bandit framework formulating a recommendation system.

In a three-party bandit framework, at each round, the learning agent filters arms and recommends a subset of arms to an arriving user, then the user makes arm selection over the recommended arms, which is observable to the learning agent as user feedback. A policy is defined as a sequential arm filtering strategy, which learns from historical user feedback and decides arm recommendation to users at each round. This three-party bandit framework is general to model a large range of user behavior, as well as a large range of arm filtering strategies.

1.1 User Behavior Modeling in Recommendation Systems

Random user preference over items. 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 and Albert, 1999; Chakrabarti et al., 2005; Ratkiewicz et al., 2010), or an item is getting increasingly more popular as it accumulates more positive feedbacks. For example, on a movie rental website, current customers tend to have more interest in Movie A that has 500 positive reviews, compared with Movie B that only has 10 positive reviews. 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 and Fristedt, 1985; Bubeck and 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.

Delayed user feedback. Customer feedbacks are often received much later than their purchasing times (e.g., a review may or may not be submitted by a customer even months later after purchasing a product). In an online learning model, policies are designed to make decisions based on historical feedback, thus, the decisions are outdated once the feedback is delayed. A realistic situation is that delays are usually random in length and could be dependent regarding item classification, in which case the collected feedback by the learner can be outdated in varying degrees, accounting for even harder estimation of items.

User click models over item ranking. When given an ordered list of recommendation, user feedback is typically regarded as their clicks on items of recommendation. Chuklin et al. (2022) has investigated such user behavior and surveyed on user click models for web search. It is shown that user click typically follows some pattern that can be modeled theoretically. Two of the most popular click models are position-based model and cascade model. Joachims et al. (2017) shows that the probability of a user examining an item depends heavily on its rank or position and typically decreases with rank. To incorporate this intuition into a click model, a set of examination parameters are defined, one for each rank, and is independent of the quality of ranked items. This position-based model (PBM) was formally introduced by Craswell et al. (2008). Another popular click model, cascade model (Craswell et al., 2008), assumes that a user scans items from top to bottom until they find a relevant item. This model allows simple estimation, since the examination events are observed: the model implies that all itmes up to the first-clicked item were examined

by a user. This also means that the cascade model can only describe sessions with one click and cannot explain non-linear examination patterns.

1.2 Arm Filtering Strategies

Arm incentivizing. In many online learning problems that utilize the MAB framework for sequential decision making (e.g., recommender systems, healthcare, finance, dynamic pricing, see Bouneffouf and 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 and Huang, 2018) that studied incentivized bandit under various settings and proposed efficient algorithms, but none of these works models practical user behavior as discussed previously.

Arm ranking. Learning to rank (L2R) is a foundational problem for recommender systems Sorokina and Cantu-Paz (2016). Solving an L2R problem amounts to understanding and predicting users browsing and clicking behaviors, so that the system can accordingly provide an optimal ranking of items to recommend to users with the aim to maximize certain rewards or utilities for the system. In the literature, L2R has been relatively well studied in the offline supervised setting, where a dataset is used to train a model in an offline fashion and then the learned model is used for ranking prediction. However, offline L2R can only provide static results that cannot

adapt to real-time data and temporal changes of the underlying ground truth. Therefore, in recent years, online L2R (ONL2R) has received increasing attention Agichtein et al. (2006). Among the existing approaches for ONL2R, the multi-armed bandit (MAB) framework is one of the most popular since it closely models the sequential interactions between the recommender system and the users (e.g., Radlinski et al. (2008); Lagrée et al. (2016); Zoghi et al. (2017); Lattimore et al. (2018)). In MAB-based ONL2R methods, the goal of the learner is to understand the click models of a spectrum of different user types through *bandit* feedback (i.e., data are collected in real-time through action-reward interactions rather than preexisting) Chuklin et al. (2015). Based on the bandit feedback, the system follows an online learning policy and *iteratively* adjusts the ranking of items to the next arriving user to maximize its long-term accumulative reward.

1.3 Overview of Major Contributions

In Chapter 3, 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 4.3). 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 and Robbins (1985). Meanwhile, the expected total incentives for both policies are upper bounded by $O(\log T)$.

In Chapter 4, We propose a new MAB model that jointly considers incentivized arm sampling, delayed sampling feedback, and self-reinforcing user preferences, all of which are important features of online recommender systems. To develop efficient and low-cost incentivized policy for this new MAB model, we propose a three-phase "UCB-Filtering-with-Delayed-Feedback" (UCB-FDF) policy, which contains an incentivized exploration phase, an incentivized exploitation phase, and a self-sustaining phase. In our UCB-FDF policy, the first two phases judiciously integrate delayed feedback information, while in third phase, the system solely relies on self-reinforcing user preferences to converge to the pulling of the optimal arm. The success of our policy design hinges upon two key insights: (i) the self-reinforcing user preference effect is actually a "blessing in disguise" and can be leveraged to establish an important "dominance" condition (more on this later) that further implies $O(\log T)$ regret and incentive costs; and (ii) the impacts of delayed feedback on regret and incentive costs can be upper bounded under appropriate statistical settings to preserve the "dominance" condition. We first show a fundamental fact that, under our UCB-FDF policy, delayed sampling feedback only has an *additive* penalty on the regret and incentive cost performances, and that this additive penalty grows logarithmically with respect to time. Specifically, we first investigate the delayed feedback impact under the assumption that the feedback delay is an i.i.d. random variable across samplings with a finite expectation. We show that the UCB-FDF policy achieves logarithmic growth rates of regret and incentive costs under this setting. Then, we

relax the i.i.d. feedback delay assumption to allow the feedback delay distribution to be arm-dependent. Under this setting, we also show that similar logarithmic growth rates of regret and incentive can still be achieved. We conduct extensive experiments on Amazon Review Data [1] to demonstrate and verify the performance of our UCB-FDF policy as well as the impacts of delayed feedback on real-world scenarios. We also verify our theoretical analysis through various product categories and demonstrate the efficacy of our proposed UCB-FDF MAB policy.

In Chapter 5, We propose the first general MAB framework that captures all key ingredients of ONL2R with position-based click models: i) two regret notions that characterize personalized and equal treatments in ranking recommendations; ii) the coupling between position preferences and mean arm rewards; and iii) partial observability of ranking position preferences. This general framework enables our rigorous policy design and analysis for MAB-based ONL2R with position-based click models. Based on the above general MAB-based ONL2R framework with position-based click models, we develop two *unified* greedy- and upper-confidence-bound (UCB)-based policies, each of which works for personalized and equal ranking treatments. For personalized treatment for ranking recommendations, we show that our greedy- and UCB-based policies achieve $\mathcal{O}(\sqrt{t}\ln t)$ and $\mathcal{O}(\sqrt{t\ln t})$ *anytime* sub-linear regrets, respectively. We show that the MAB policy design for the equal treatment case is more challenging, which may require solving an NP-hard problem in each time step depending on the collective utility function for social welfare. To address this challenge, we identify classes of utility functions and establish their associated sufficient conditions of approximation accuracy, under which $\mathcal{O}(\sqrt{t}\ln t)$ and $\mathcal{O}(\sqrt{t\ln t})$ *anytime* sublinear regrets are still achievable for greedy- and UCB-based policies, respectively.

[1] https://nijianmo.github.io/amazon/

Chapter 2

Related Work

Multi-armed bandit have seen an increasing interest in the academia and industry over the last few years. It is a special area of reinforcement learning that is specialized in making decisions under uncertainty. In recommendation systems, multi-armed bandit is a powerful tool to model online user-platform interactions, which allows a large range of real-world user behavior, e.g., random user preference over items, delayed user feedback, and customized user click models over an ordered list of items. In this chapter, we provide an overview of some closely related fields with our work.

2.1 Bandits with Random User Preferences

The impacts of random user preferences in e-commerce platforms have received increasing interest in several different areas in learning and economics. Existing works in (Agrawal et al., 2017, 2019) formulated the user preference variation given different product bundles by the multi-nomial logit model on top of the bandit learning framework and proposed a Thompson Sampling approach that achieves a worst-case regret bound of $O(\sqrt{NT}\log TK)$, where N is the size of recommended arm bundle. With a different focus on preference modeling, Barabási and Albert (1999); Chakrabarti et al. (2006); Ratkiewicz et al. (2010) investigated the network evolution with "preferential attachment" that formulates the social behavior known as self-reinforcing preferences.

Also, a similar social behavior, referred to as *herding*, is studied in the Bayesian learning model literature (Bikhchandani et al., 1992; Smith and Sørensen, 2000; Acemoglu et al., 2011). For example, Acemoglu et al. (2011) first studied the conditions under which there exists a convergence in probability to the desired action as the size of a social network increases. More recently, Shah et al. (2018) incorporated positive externalities in user arrivals and proposed 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.

2.2 Bandits with Delayed User Feedback

Motivated by practical issues in the clinical trials, Eick (1988) was the first to introduce a two-armed bandit model with delayed responses, where the patients survival time reports after the treatment are delayed. Recently, Joulani et al. (2013) provided a systematic study and showed that for delay τ with a finite expectation, the worst case regret scales with $O(\sqrt{KT\log T} + K\mathbb{E}[\tau])$, where K is the number of arms. Meanwhile, Vernade et al. (2017) showed that stochastic MAB problems with delayed feedback have a regret lower bound $O(K\log T)$. However, this work assumed that the distribution of the random delay is arm-independent. In contrast, Joulani et al. (2013) considered arm-dependent delay distributions that have an upper bound of the maximum random delay. More recently, Manegueu et al. (2020) considered

arm-dependent and heavy-tailed delay distributions, where only an upper bound on the tail of the delay distribution is needed, without requiring the expectation to be finite. Also, Lancewicki et al. (2021) studied the case where the delay distribution is reward-dependent, which implies that the random delay in each round may also depend on the reward received on the same round. However, most of these works on delayed bandits are based on the standard stochastic MAB framework. In contrast, we consider delayed feedback in incentivized bandit learning with self-reinforcing user preferences, which is a more appropriate model for real-world recommender systems than the standard stochastic MAB.

2.3 Arm Filtering Strategies

In a classic bandit model, an action is defined as pulling one arm from available arm set. When applied to recommendation systems, bandit model often formulates the recommendation as arm pulling, and pulling one arm implies recommending one item to users, which is too naive in real world. A more general formulation is defining actions taken by the learner as an arm filtering strategy, which allows more than one arms recommended to users, and allows more flexibility of user behavior formulation. Here, we focus on two arm filtering strategies: arm incentivizing and arm ranking.

Arm incentivizing. incentivized MAB has attracted growing attention in recent years (Kremer et al., 2014; Frazier et al., 2014; Mansour et al., 2015, 2016; Wang and Huang, 2018). To our knowledge, Frazier et al. (2014) first adopted incentive schemes into a Bayesian MAB setting. In their model, the 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 and Huang (2018) considered the non-Bayesian setting with non-discounted rewards. Agrawal and Tulabandhula (2020) considered incentivizing exploration under contextual bandits. These models differ from ours in both the incentive schemes and user behaviors.

Arm ranking. research on MAB-based online learning to rank (ONL2R) remains in its infancy. To our knowledge, the first work that studied MAB-based ONL2R was reported in Radlinski et al. (2008), which, however, is based on the cascade click models. Later, MAB-based ONL2R with position-based click model was considered in Lagrée et al. (2016), which established an $\Omega(\log T)$ regret lower bound for their model and proposed algorithms with a matching regret upper bound. Although sharing some similarity to ours, the position-based click model in Lagrée et al. (2016) is a simpler model setting, which assumes *known* position preference. In contrast, the unknown position preferences in our work requires extra learning besides conventional arm mean estimation, causing non-trivial policy design and performance trade-off. Generalized click model encompassing both position-based and cascade click models was proposed in Zoghi et al. (2017), where the authors also developed a BatchRank policy with a gap-dependent upper bound on the T-step regret.BatchRank was later outperformed by the TopRank policy proposed in Lattimore et al. (2018) in both cascade and position-based click models. We note that all these existing works make a strong and unrealistic assumption that user behavior is *homogeneous*, and they all aim to optimize a standard MAB objective, i.e., maximizing total clicks. Similar to our model, combinatorial semi-bandit (CSB) also considers multiple arms being

pulled with semi-bandit feedback at each round (Kveton et al., 2015; Chen et al., 2016b,a; Wang and Chen, 2018). For example, the work in Chen et al. (2016a) studied the general CSB framework, and proposed a UCB-style policy CUCB with regret upper bound. Later, the work in Wang and Chen (2018) developed a policy based on Thompson sampling. Also, the work in Kveton et al. (2015) derived two upper bounds on the n-step regret of policy CombUCB1, while proving a matching lower bound using a partition matroid bandit. We note that the CSB setting differs from ours in two key aspects: i) while CSB considers a subset of arms as a super arm at each round, our setting additionally considers the ranking *within* the super arm that also affects the reward; ii) the reward in our setting is based on *two unknown parameters*: position preferences and arm means, which cannot be directly estimated separately due to partial observation of the user feedback. These complications are unseen in conventional CSB settings.

Chapter 3

Bandit Learning with Self-Reinforcing User Preferences

3.1 Overview

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 work. 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 recommendation systems. This is in stark contrast to most existing works in the areas of incentivized bandits (Frazier et al., 2014; Wang and Huang, 2018), where a (unrealistic) deterministic greedy user behavior is often assumed. Under 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.

3.2 System Model and Problem Formulation

We denote the set of arms offered by the agent as $\mathcal{A} = \{1, \ldots, M\}$. Each arm a follows a Bernoulli reward distribution P_a with an unknown mean $\mu_a > 0$. The process runs for T rounds. As shown in Fig. 5.2, in each time step $t \in \{1, \ldots, T\}$, a user arrives and chooses an arm $I(t)$ to pull, then receives a random reward $X(t) \sim P_{I(t)}$, which is observable to the agent. We use $T_a(t) = \sum_{i=1}^{t} 1_{\{I(i)=a\}}$ 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 $S_a(t) = \sum_{i=1}^{t} X(i) \cdot 1_{\{I(i)=a\}}$. We let $T_a(0) = 0$ and $S_a(0) = 0$, $\forall a \in \mathcal{A}$. We assume that there is a unique best arm $a^* \in \mathcal{A}$, 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 $\sum_{a\in\mathcal{A}} \lambda_a(t) = 1, \forall t$. 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:

$$\lambda_a(t) = \frac{F\big(S_a(t-1) + \theta_a\big)}{\sum_{i\in\mathcal{A}} F\big(S_i(t-1) + \theta_i\big)}, \tag{3.1}$$

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.

Several important remarks for the preference model in (3.1) 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 and Little, 2008; Gupta, 1988). The multinomial logit model is also used in the social network literature to model preferential attachment (Barabási and Albert, 1999), where the probability that a link connects a new node j with another existing node i is linearly proportional to the degree of i. Notably, this multinomial logit model has also been adopted in Shah et al. (2018) to model the same type of self-reinforcing

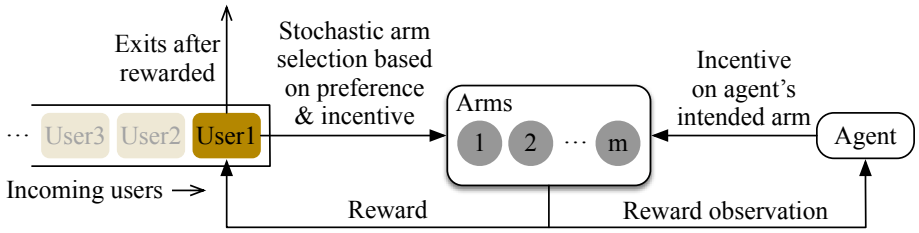


Figure 3.1: Incentivized MAB model with stochastic arm selection based on user preference rates and incentives.

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. 5.2). 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 and Krishnamurthi, 1996; Bawa and 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 and 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

[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.

rates of the arms with incentive b are updated as follows:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{\bar{G}(b,t) + F\big(S_i(t-1)+\theta_i\big)}{\bar{G}(b,t) + \sum_{j\in\mathcal{A}} F\big(S_j(t-1)+\theta_j\big)}, & i = a, \\ \dfrac{F\big(S_i(t-1)+\theta_i\big)}{\bar{G}(b,t) + \sum_{j\in\mathcal{A}} F\big(S_j(t-1)+\theta_j\big)}, & i \neq a, \end{cases} \tag{3.2}$$

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 $\sum_{i\in\mathcal{A}} F\big(S_i(t-1)+\theta_i\big)$ 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 (3.2) still follows the multinomial logit model. Also, we can see from (3.2) 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 $\sum_{i\in\mathcal{A}} F\big(S_i(t-1)+\theta_i\big)$ and let $G(b,t) \triangleq \bar{G}(b,t)/\sum_{i\in\mathcal{A}} F\big(S_i(t-1)+\theta_i\big)$. Then, it can be verified that Eq. (3.2) 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 $B_t \coloneqq \sum_{i=1}^{t} b_t$, 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 $\mu_{a^*} T$.

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.

3.3 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.

3.3.1 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 work 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.

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 and 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 and 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, 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.

3.3.2 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 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:

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

Algorithm 1 At-Least-n Explore-Then-Commit (ALnETC).

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

1: **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: **2) Exploitation Phase:** Incentivize pulling the best-empirical arm $\hat{a}^* \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a(\tau_n)$ 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: **3) Self-Sustaining Phase:** Users pull arms based on their own preferences until time T.

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 $G(b,t) \triangleq \bar{G}(b,t) / \sum_{i \in \mathcal{A}} F\big(S_i(t-1) + \theta_i\big)$ (cf. Section 4.2). 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" $\sum_{i \in \mathcal{A}} F(S_i(t-1) + \theta_i)$ 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) $q \geq (2 \max_{a \neq a^*} \mu_a)/\Delta_{min}^2$, (iii) $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then the expected regret of ALnETC is upper bounded by:*

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

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 $q < (2 \max_{a \neq a^*} \mu_a)/\Delta_{min}^2$, 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

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

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 $\gamma \in (0, 1/4)$. □

3.3.3 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:

Algorithm 2 UCB-List

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

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

2: **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.

3: **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.

4: **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 5. 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 $a \in \mathcal{A}, a \neq a^*$ 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. □

3.4 Simulations

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

3.4.1 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

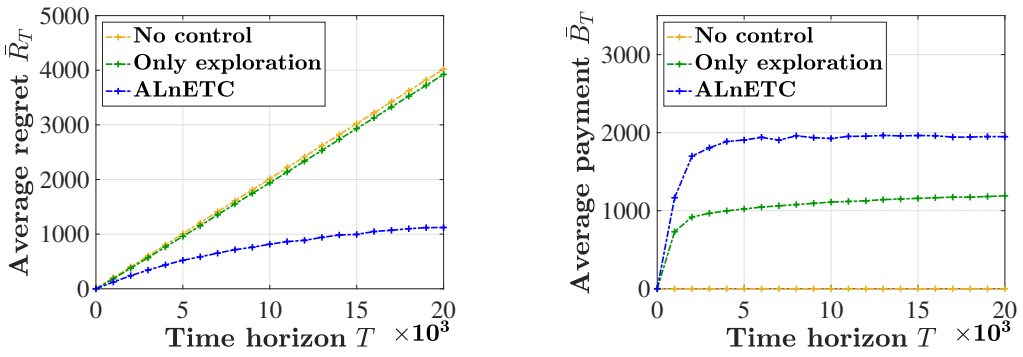
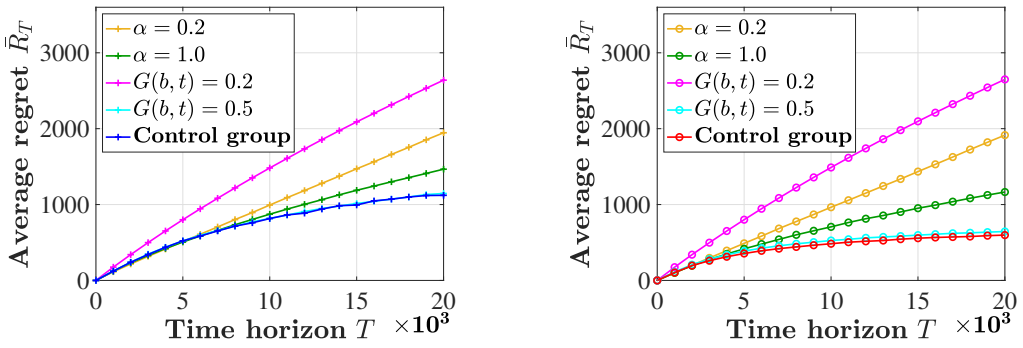


Figure 3.2: Comparison of ALnETC and baselines.

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. 3.2, 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.

3.4.2 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 3.4.1: a two-armed model with means $\boldsymbol{\mu} = [0.3, 0.5]$ and initial biases $\boldsymbol{\theta} = [100, 1]$, the feedback function



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

Figure 3.3: Comparisons of imperfect conditions.

$F(x) = x^\alpha$. 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. 3.3, 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.

3.4.3 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. 4.5-3.7, where each data point is averaged over 1000 trials. Fig. 4.5 illustrates the performance of both average regret and total payment. Fig. 4.5 also serves as a benchmark for comparisons with other three groups of results. In each of Figs. 4.6–3.7, only one parameter is changed compared to the benchmark group. This helps us observe the changes in average regret and total payment. In Fig. 4.6, all settings are the same as Fig. 4.5 except $\alpha = 2$. In Fig. 4.7, all settings are the same as those in Fig. 4.5 except $\boldsymbol{\theta} = [50, 50, 1]$. In Fig. 3.7, all settings are the same as Fig. 4.5 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. 4.6, 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. 4.7, both policies have larger average regrets because it takes more effort for both policies to mitigate the larger initial biases. In Fig. 3.7, 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.

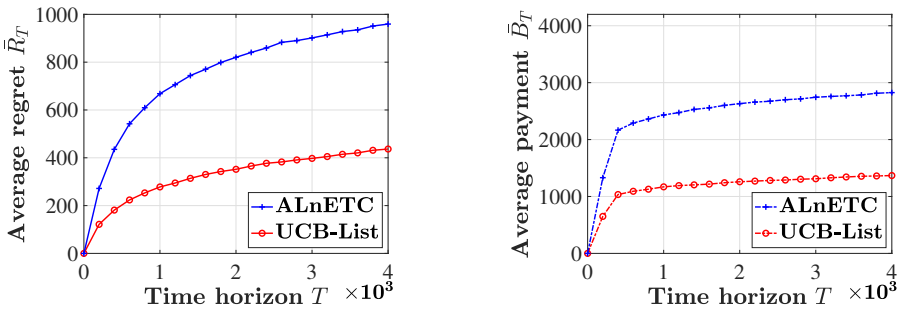


Figure 3.4: Benchmark results.

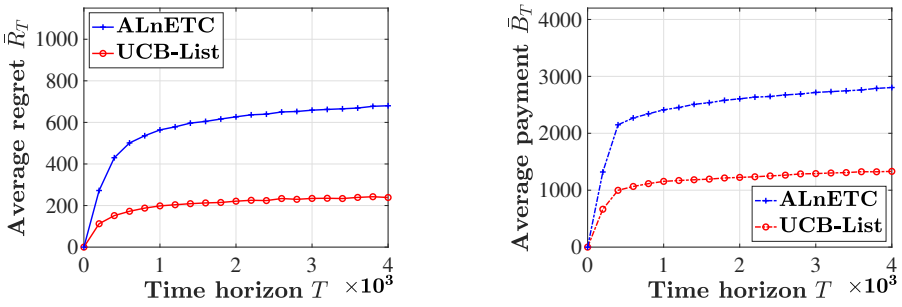


Figure 3.5: Policy performance with parameter $\alpha = 2$.

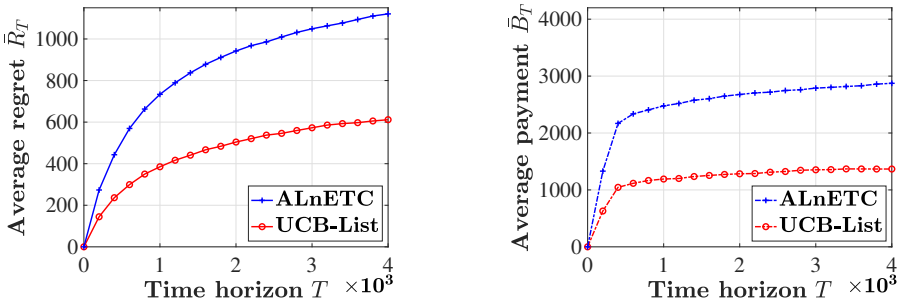


Figure 3.6: Policy performance with parameter $\boldsymbol{\theta} = [50, 50, 1]$.

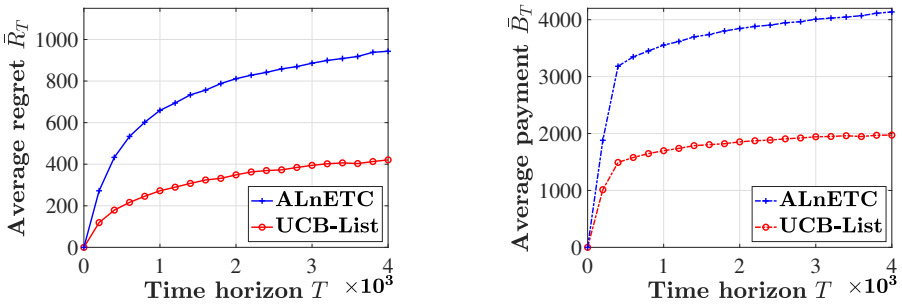


Figure 3.7: Policy performance with parameter $b = 1.8$.

Chapter 4

Bandit Learning with Joint Effect of Incentivized Sampling, Delayed Sampling Feedback, and Self-Reinforcing User Preferences

4.1 Overview

It is worth noting that most of the existing MAB models in the literature have not considered the joint effect of *three* common phenomena in e-commerce recommendation systems: (i) In many e-commerce recommendation systems, the platform (the learning agent) cannot sample an intended product (an intended arm) directly and has to incentivize customers (e.g., through promotions and coupons) to sample the product and receive the sampling feedback from the customers indirectly (e.g., ratings and reviews); (ii) Customer feedbacks are often received much later than their purchasing times (e.g., a review may or may not be submitted by a customer even months later after purchasing a product); and (iii) Customer preferences among products are influenced and reinforced by historical feedbacks, which may even lead to various viral effects over some products (the more good reviews one product has received, the more likely that the next arriving customer will prefer this product). The lack of a fundamental understanding and joint studies of these three important factors in MAB policy designs motivates us to fill this gap in this work.

Toward this end, we propose a new MAB framework that *jointly* considers i) incentivized sampling, ii) delayed sampling feedback, and iii) self-reinforcing user preferences in online recommendation systems. However, we note that the MAB policy design for the proposed new MAB framework is highly non-trivial due to the complex couplings between the aforementioned three factors. First, similar to conventional MAB problems, there exists a dilemma between sufficient *exploration* through sampling to learn an optimal arm (i.e., an optimal product), which may incur numerous pullings of sub-optimal arms, and the greedy *exploitation* to play the arm that has performed well thus far to earn profits. Second, there is another dilemma to the learning agent between offering sufficiently attractive incentives to mitigate biases (due to lack of initial data and self-reinforcing user preferences) and avoid spending unnecessarily high incentives that hurt the learning agent's profits. Last but not least, the delayed sampling feedbacks may render the estimation of arms' quality during the MAB process highly inaccurate, introducing yet another layer of uncertainty to the MAB online learning problem, which is already plagued by complications from incentivized sampling and self-reinforcing user preferences. As in most MAB problems, we adopt "regret" as our performance metric in this work, which is defined as the cumulative reward gap between the proposed policy and an optimal policy design in hindsight. Under the regret setting, the complications due to these three key factors naturally prompt the following fundamental questions:

(1) How should the agent design an incentivizing strategy to strike a good balance between exploration and exploitation to achieve sublinear (hopefully logarithmic) regrets?

(2) To avoid offering exceedingly high incentives, how should the agent incentivize

in order to attract a user crowd that prefer an optimal arm, so that the users' self-reinforcing preference could automatically gravitate toward this optimal arm without further incentives?

(3) Under various delayed feedback situations in the new MAB framework (e.g., unbounded random delays, heavy-tailed delay distributions, and arm-dependent delays), could we still achieve low regrets with low incentive costs?

In this chapter, we answer the above fundamental questions affirmatively by proposing a new "Delayed-UCB-Filtering" policy for the MAB framework that jointly considers incentivizing sampling, delayed sampling feedback, and self-reinforcing user preferences. We show that our proposed policy achieves $O(\log T)$ regret with $O(\log T)$ incentive payments. The success of our policy design hinges upon two key insights: (i) the self-reinforcing user preference effect is actually a "blessing in disguise" and can be leveraged to establish an important "dominance" condition (more on this later) that further implies $O(\log T)$ regret and incentive costs; and (ii) the impacts of delayed feedback on regret and incentive costs can be upper bounded under appropriate statistical settings to preserve the "dominance" condition.

4.2 System Model and Problem Formulation

The system has a set of $M \geq 2$ arms denoted by $\mathcal{A} = \{1, \ldots, M\}$, and each arm a follows a Bernoulli reward distribution P_a with an unknown mean $\mu_a > 0$. The bandit time horizon has T rounds. In each time step $t = 1, 2, \ldots, T$, a user arrives and chooses an arm $I(t)$ to pull. Then, the user will receive a random reward feedback $X(t) \sim P_{I_t}$. Both the arm selection $I(t)$ and the feedback $X(t)$ are 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 arm a is

pulled up to time step t. We let $T_a(0) = 0$, $\forall a \in \mathcal{A}$. We assume that there is a unique best arm $a^* \in \mathcal{A}$ in the sense that $a^* = \arg\max_{a \in \mathcal{A}} \mu_a$ and let $\mu^* = \mu_{a^*}$. Also, we define $\Delta_a \triangleq \mu^* - \mu_a$ as the gap between the mean of the optimal arm and the mean of arm a.

4.2.1 Delayed Feedback Modeling

In this work, we consider delayed feedback, i.e., when an arm $I(t)$ is pulled at time step t, the corresponding Bernoulli reward $X(t)$ is observed after a delay period $\tau_{I(t),t}$, i.e., the feedback $X(t)$ is observed at time step $t + \tau_{I(t),t}$. Without loss of generality, we model the random delay time as a random variable $\tau_{a,t} \sim \mathcal{T}_a$, where the delay distribution $\mathcal{T}_a$ of arm a is unknown to the agent.

We consider two settings of delayed feedback. We first consider i.i.d. delays $\{\tau_t\}_{t \leq T}$ across time and arms, i.e., the delay distributions are identical for all arms. Thus, we omit the arm index in the notations of delay feedback in this setting. Next, we generalize the delay modeling by allowing arm-dependent delay distributions, where the delay distributions are allowed to differ across arms. In both settings, we do not make further assumptions on the delay distributions, except that we only require a finite delay expectation. Note that we allow the support of the delays to be unbounded, i.e., an infinite delay time is possible in both settings. This models the practical scenarios in online recommendation systems that some user feedbacks (e.g., ratings and reviews) may never be received.

Under delayed feedbacks, we denote the total number of missing feedbacks from arm a up to a time step t as $D_a(t) \triangleq \sum_{s=1}^{t} 1_{\{s+\tau_{a,s}>t\}}$. We let $D_a^*(t) = \max_{1 \leq s \leq t} D_a(s)$, $\forall a \in \mathcal{A}$ as the maximum total number of delayed feedback for arm a up to time

t. Note that $D_a^*(t) = 0$, $\forall a \in \mathcal{A}$ corresponds to the non-delayed setting. In this case, $T_a(t)$ denotes the total number of pulling times of arm a up to time t. At each time step t, the agent observes a set of time-stamped feedback denoted by $S_t \subset \mathbb{N} \times \{0, 1\}$. In the set S_t, each element is a pair of time index and a Bernoulli reward value, and the time index is the time step when the corresponding reward is observable. Note that in this model, by observing the set S_t, the agent is aware of the information of both the time step when the feedback is received, and the arm that generated the feedback. We denote the total reward generated by arm a up to time t as $S_a(t) \triangleq \sum_{s=1}^{t} X(s) \cdot 1_{\{I(s)=a, s+\tau_{a,s} \leq t\}}$, and let $S_a(0) = 0$, $\forall a \in \mathcal{A}$.

4.2.2 User Preferences and Incentive Impact Modeling

In this work, we assume that the arrival at time t has a non-zero probability $\lambda_a(t) \in (0, 1)$ to pull each arm $a \in A$. We note that $\lambda_a(t)$ can also be thought of as the user's preference rate of arm a, and $\sum_{a \in \mathcal{A}} \lambda_a(t) = 1$, $\forall t \leq T$. We adopt the widely accepted multinomial logit model in the economics literature(Bawa and Shoemaker, 1987) to model arm a's preference rate at time step t as follows:

$$\lambda_a(t) = \frac{F\big(S_a(t-1) + \theta_a\big)}{\sum_{i \in \mathcal{A}} F\big(S_i(t-1) + \theta_i\big)}, \tag{4.1}$$

where $F(\cdot) : \mathbb{R} \to (0, +\infty)$ is a feedback function that is increasing, and $\theta_a > 0$ denotes a fixed initial preference bias of arm a. We note that the preference rate modeling in (Zhou et al., 2021) is also based on the multinomial logit model, which appears to be in the same form as in (4.1). However, the key difference between our preference model in (4.1) and that in (Zhou et al., 2021) is that the accumulative

award information $S_i(t-1)$ in (4.1) accounts for reward information that can only be observed up to time t. In other words, $S_i(t-1)$ in (4.1) is affected by feedback delays. In fact, the preference model in (Zhou et al., 2021) can be viewed as a special case of our model with zero delay.

Since the arriving users select arms based on preferences, while the agent aims to maximize the total reward in the long run, there exists a general difference between users' arm preferences and agent's intended arm selection. To induce users to pull arms following the agent's goal, the agent needs to intervene users' arm pulling by offering incentives on its desired arm, so as to increase the user preference of pulling the arm. That is, the agent incentivizes arm $I'(t)$ at time step t so that $\lambda_{I'(t)}(t)$ increases accordingly. Note that when $\lambda_{I'(t)}(t)$ increases, the preference rates on the other arms will decrease since $\sum_{a\in\mathcal{A}} \lambda_a(t) = 1$, $t \leq T$. We adopt the "coupon effect" model, which is widely used in the economics and marketing literature (Bawa and Shoemaker, 1987). Specifically, we consider a fixed incentive b in each time step and denote the time-dependent incentive impact as $G(b,t)$. Then, the posterior preference rates of the arms with incentive b are updated as follows:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{\lambda_i(t) + G(b,t)}{1 + G(b,t)}, & i = a, \\ \dfrac{\lambda_i(t)}{1 + G(b,t)}, & i \neq a. \end{cases} \tag{4.2}$$

We remark that the definition of the posterior preference update in (4.2) also follows from the multi-nomial logit model, which is widely used to model user preferences and their variations in bandit field (Chen and Wang, 2017; Avadhanula, 2019; Dong et al., 2020; Zhou et al., 2021). Based on the defined posterior preference, as incentive

impact $G(b,t)$ increases to infinity (either the incentive value b increases to infinity or the users are more sensitive to incentives as time goes by), the user preference will be induced to pulling the agent's desired arm a with probability one. For further detailed interpretations of the incentive impact function $G(b,t)$, we refer readers to the literature (e.g., (Zhou et al., 2021)). Note also that, due to the random user behaviors, it is possible that $I'(t) \neq I(t)$, i.e., the arm that the agent incentivizes is not the one that a user pulls eventually. We define the accumulative incentive up to time step t as $B_t \triangleq \sum_{s=1}^{t} b_t$, where $b_t \in \{0, b\}$, $\forall t \leq T$, denotes the agent's binary decision whether to offer incentive b at time step t.

4.2.3 Regret Modeling

As in most bandit learning problems, the goal of the agent is to maximize the total expected reward $\mathbb{E}\big[\sum_{a\in\mathcal{A}} S_a(T)\big]$ in the long run. Toward this end, we need the notion of the oracle incentivized policy, where in hindsight, the agent is aware of the optimal arm a^* and can always offer an infinite amount of payments to users with feedback being observable immediately, so that the posterior preference rate of arm a^* is always infinitely close to one. As a result, the expected accumulative reward generated under the oracle policy up to time T is $\mathbb{E}[S_{a^*}(T)] = \mu^* \cdot T$. However, since the optimal arm a^* is unknown to the agent, the goal of the agent is to maximize the total expected reward $\mathbb{E}[\Gamma_T]$ in the long run by designing an incentivized policy with low accumulative incentive in the presence of self-reinforcing preferences and feedback delay. Similar to conventional MAB, we measure the performance gap between our accumulative reward against that of the oracle policy, which is denoted by regret R_T.

The expected (pseudo) regret is defined as follows:

$$\mathbb{E}[R_T] = \mu^* \cdot T - \mathbb{E}\big[\sum_{a \in \mathcal{A}} S_a(T)\big].$$

Our goal is to minimize $\mathbb{E}[R_T]$ with low expected accumulative payment $\mathbb{E}[B_T]$, i.e., sub-linear growth rate regarding time horizon T. It is clear that any policy with bounded payment cannot outperform the oracle policy. Thus any expected regret defined by comparing with bounded-payment policy is upper bounded by our regret.

4.3 Policy Designs and Performance Analysis

In this section, we first present the general version of the UCB-FDF policy that works with any delay distributions, where we upper bound the delay impact on the regret and incentive costs. Based on this general result, we then study the regret and incentive costs performance of UCB-FDF under the assumptions of 1) i.i.d. feedback delay across arms/times and 2) arm-dependent delay distributions. In both cases, we denote the total number of missing feedbacks over all arms by $D(t) \triangleq \sum_{a \in \mathcal{A}} D_a(t)$, and denote the maximum number of missing feedbacks during the first t time steps by $D^*(t) \triangleq \max_{1 \le s \le t} D(s)$. For arm a at time step t, we denote the number of its pulling times whose feedback is observed by $T'_a(t) = T_a(t) - D_a(t)$, and denote the maximum mean gap by $\Delta^* = \max_{a \in \mathcal{A}} \Delta_a$. At time step t, we denote the sample mean estimation (due to delayed feedbacks) of arm a by $\hat{\mu}_a(t) = S_a(t)/T'_a(t)$. Our UCB-FDF policy is illustrated in Algorithm 3.

UCB-FDF policy contains three phases: an incentivized exploration phase, an incentivized exploitation phase, and a self-sustaining phase. UCB-FDF policy tack-

Algorithm 3 The UCB-Filtering-with-Delayed-Feedback Policy (UCB-FDF).

Require: Time horizon T and incentive payment b, the confidence interval of arm a at time step t defined as $c_a(t) = \sqrt{\ln T / \big(2T'_a(t)\big)}$.

1: **Initialization:** Incentivize pulling the arms satisfying $T'_a(t) = 0$ with incentive payment b until $\min_{a\in\mathcal{A}} T'_a(t) \geq 1$. Let set $\mathcal{U} = \mathcal{A}$. Mark current time as t_0.

2: **Exploration Phase:** While $|\mathcal{U}| > 1$, remove all the arms from set $\mathcal{U}$ satisfying $\hat{\mu}_a(t) + c_a(t) \leq \max_{i\neq a, i\in\mathcal{U}} \big(\hat{\mu}_i(t) - c_i(t)\big)$ if there is any, then incentivize pulling arm $a \in \arg\min_{i\in\mathcal{U}} T'_a(t)$ with payment b. If $|\mathcal{U}| = 1$, let arm $\hat{a}^* = \{a : a \in \mathcal{U}\}$ and mark current time as t_1.

3: **Exploitation Phase:** Incentivize pulling arm $\hat{a}^*$ with payment b until it dominates: $S_{\hat{a}^*}(t) \geq \sum_{a\neq\hat{a}^*} \big(S_a(t) + D_a(t)\big)$. Mark current time as t_2.

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

les feedback delays in the following two key aspects: (i) correcting the sample mean estimate of arms by only considering the number of pulling times that have observed feedback, (ii) setting the length of the exploitation phase in such a way that the outstanding rewards do not harm the emergence of "dominance" (i.e., one arm receiving at least half of the rewards) of the sampled optimal arm. Subsequently, these two aspects also influence the regret and incentive. In order to have enough arm exploration with an unbiased sample mean estimate, the loss of counted number of pulling times necessitates a carefully designed exploration phase that incentivizes the pulling of the least informed arm $a \in \arg\min_{i\in\mathcal{U}} T'_a(t)$ under delayed feedbacks. Similarly, the delay-based dominance threshold (i.e., $S_{\hat{a}^*}(t) \geq \sum_{a\neq\hat{a}^*} \big(S_a(t) + D_a(t)\big)$ in Step 3 of Algorithm 3) guarantees the dominance of sampled optimal arm, while also accounts for a longer exploitation phase to mitigate the delayed feedback effect. We now analyze the upper bounds of the pseudo regret and expected incentive of the UCB-FDF policy.

Lemma 5. (UCB-Filtering-with-Delayed-Feedback) *Given a fixed time horizon T, if*

$G(b,t) > 1$, and $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$[1], then the pseudo regret of Algorithm 3 $\mathbb{E}[R_T]$ is upper bounded by

$$\mathbb{E}[R_T] \leq \sum_{a \neq a^*} \frac{8\Delta_a\big(G(b,1)-1\big) + 8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2} \ln T + \frac{g(F,1)\Delta^*\big(\mathbb{E}[D^*(T)] + 4K\big)}{g(b,1)-1},$$

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

$$\mathbb{E}[B_T] \leq b \cdot \frac{2G(b,1)+1}{G(b,1)-1}\left[\frac{8\ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \frac{8\ln T}{\Delta_a^2} + \mathbb{E}[D^*(T)] + 4K\right].$$

Remark 7. The UCB-FDF policy achieves a sub-linear total incentive cost by leveraging the property of self-reinforcing preference. We can show that as long as the self-reinforcing preference function $F(x)$ satisfies the condition $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$, then "monopoly" happens with probability one (i.e., the scenario where only one arm has positive probability to be pulled, thus this particular arm is the only preferred arm). A natural incentivizing policy is to incentivize sampled optimal arm until it achieves monopoly. However, the key challenge here is that the onset of monopoly could take infinite time steps, which implies linear total incentive. Moreover, self-reinforcing property is not merely disrupting the system from converging to the optimal arm. The key idea in our UCB-FDF policy design is that under the condition of the self-reinforcing preference function $F(\cdot)$, after one arm establishes its dominance (i.e., the arm a generates at least half of the current total reward), it will have exponentially increasing probability to beat other arms and achieve monopoly. More importantly, we can show that the onset of arm dominance takes sub-linear

[1]The notation $\Theta()$ in this work is defined as that, if $F(x) = \Theta(g(x))$, then there exist x_0 and two constants $C_1, C_2 > 0$, such that $C_1 G(x) \leq F(x) \leq C_2 g(x)$ for all $x \geq x_0$.

times, thus allowing us to achieve sub-linear total incentive costs.

Remark 8. The feedback delay affects the observation of arm dominance, since the missing reward information from suboptimal arms, if not compensated carefully, can potentially destroy the dominance status of the optimal arm. Thus, to guarantee dominance of the optimal arm, a longer exploitation phase is necessary, and thus a large total incentive is required.

The existence of delays in our MAB model introduces an additive term $\Theta(\mathbb{E}[D^*(T)])$ in both regret and incentive costs, which is dependent on the maximum accumulated delayed feedback up to time horizon T. Based on Lemma 5, in what follows, we will analyze the upper bounds of the expected maximum accumulated delayed feedback under different assumptions on delay distributions.

4.3.1 Arm-Independent Delay with a Finite Expectation

We now analyze the delay impact under our first assumption. In the arm-independent case, we consider an i.i.d. sequence $\{\tau_t\}_t$ of random delay regarding time step $t \leq T$. We do not make any assumption on the shape of the delay distribution, except that we only assume a finite expectation $\mathbb{E}[\tau_1]$. Thus, an infinite random delay is possible under this assumption, implying some feedbacks may never be observed by the agent. Our results show that under this assumption, we can still achieve similar orders of the regret and incentive costs growth rates, since the key fact is that we can upper bound the expected number of such unexpectedly large random delays for every time step t.

Existing works (e.g., (Joulani et al., 2013)) provided a systematic study on the delay effect on the partial monitoring problem with side information, including the

stochastic problems. Although these works only considered the classic stochastic MAB, they share some similarities with our work in that their analysis of delay effects also leveraged the maximum number of missing feedbacks during the first t time steps $D^*(t)$. However, since our UCB-FDF policy has a different structure compared to these works on delayed stochastic MAB, their delay analysis is not applicable to our policy. Next, we restate a result in (Joulani et al., 2013), which will be useful in our analysis.

Lemma 6 (Lemma 2 in Joulani et al. (2013))**.** *Assume* $\{\tau_1, \ldots, \tau_t\}$ *is a sequence of i.i.d. random variables with finite expected value, and let* $B(t, s) = s + 2\log t + \sqrt{4s \log t}$*. Then, it holds that*

$$\mathbb{E}[D^*(t)] \leq B(t, \mathbb{E}[\tau_1]) + 1.$$

Theorem 7. (Arm-Independent Delay) *Under i.i.d. delays with a finite expectation and the conditions of Lemma 5, the pseudo regret of Algorithm 3* $\mathbb{E}[R_T]$ *is upper bounded by*

$$\left[\frac{2G(b,1)\Delta^*}{G(b,1)-1} + \sum_{a \neq a^*} \frac{8\Delta_a\big(G(b,1)-1\big) + 8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2}\right] \ln T + \frac{G(b,1)\Delta^*\big(\sqrt{4\mathbb{E}[\tau_1]\ln T} + \mathbb{E}[\tau_1] + 4K + 1\big)}{G(b,1)-1},$$

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

$$b \cdot \frac{2G(b,1)+1}{G(b,1)-1}\left[\left(2 + \frac{8}{\Delta_{min}^2} + \sum_{a \neq a^*} \frac{8}{\Delta_a^2}\right)\ln T + \sqrt{4\mathbb{E}[\tau_1]\ln T} + \mathbb{E}[\tau_1] + 4K + 1\right].$$

We note that the gap summation of arms $\sum_{a \neq a^*} \Delta_a$ plays an important role in both regret and total incentive. As arm gaps getting smaller, it is more difficult to distinguish the optimal arm from others. Thus, a longer exploration phase is required to conduct enough sampling, which implies a larger regret and a larger total incentive costs. On the other hand, the feedback delay causes additive terms in both regret and incentive costs in terms of the expected delay $\mathbb{E}[\tau_1]$, and the delay impact can be upper bounded as long as the expected delay $\mathbb{E}[\tau_1]$ is no larger than time horizon T.

4.3.2 Arm-Dependent Delay with Finite Expectations

Now, we further relax the assumption on the delay to allow *arm-dependent* delays. In this case, the delay has two key impacts on the system: (i) for each arm, there is a different real-time information loss when estimating the sample mean, (ii) for the whole arm set, different scales of delay cause an *uneven* arm estimation, which results in a larger risk of the elimination of the optimal arm in the UCB-based exploration step. We formally state our arm-dependent delay assumption as follows:

Assumption 1. *The delays of arm $a \in \mathcal{A}$ form an independent delay sequence $\{\tau_{a,t}\}$, where each element is a random variable satisfying $\tau_{a,t} \sim \mathcal{T}_a$, with a finite expectation $\mathbb{E}[\tau_{a,1}] < +\infty$, $\forall a \in \mathcal{A}$.*

Under Assumption 1, we show a more general result on the upper bound of $\mathbb{E}[D^*(t)]$ as follows:

Lemma 8. *Under Assumption 1, given a finite number of arms $K > 0$, it holds that*

$$\mathbb{E}[D^*(t)] \leq \sum_{a \in \mathcal{A}} 2\mathbb{E}[\tau_{a,1}] + 3K \log \frac{t}{K}.$$

The result in Lemma 8 implies larger upper bounds of the regret and incentive, due to the existence of the pre-log factor K. This is a consequence of the situation where, as we consider arm-dependent delay distributions, the worst case could be evenly distributed expected delays $\mathbb{E}[\tau_{a,1}]$ of arm a with respect to time horizon T. Formally, we state the upper bounds of regret and incentive as follows:

Theorem 9. (Arm-Dependent Delay) *Under Assumption 1 and the conditions of Lemma 5, the pseudo regret of Algorithm 3* $\mathbb{E}[R_T]$ *is upper bounded by*

$$\sum_{a\neq a^*} \frac{8\Delta_a\big(G(b,1)-1\big)+8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2}\ln T + \frac{G(b,1)\Delta^*\big(3K\ln\frac{T}{K}+\sum_{a\in\mathcal{A}}2\mathbb{E}[\tau_{a,1}]+4K\big)}{G(b,1)-1},$$

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

$$b\cdot\frac{2G(b,1)+1}{G(b,1)-1}\left[\left(\frac{8}{\Delta_{min}^2}+\sum_{a\neq a^*}\frac{8}{\Delta_a^2}\right)\ln T + 3K\ln\frac{T}{K}+\sum_{a\in\mathcal{A}}\mathbb{E}[\tau_{a,1}]+4K\right].$$

Similar to the results under the i.i.d. delay assumption, we can still upper bound the regret and incentive by an logarithmic growth rate $O(\log T)$ under arm-dependent delay. This implies that even under the weak delay assumption where only finite expectation is needed, UCB-FDF can estimate arms without too much bias, and finally achieve logarithmic regret with logarithmic incentive costs.

4.4 Experiments

In this section, we first introduce our experiment setting and the dataset, then illustrate our experimental results.

4.4.1 Experimental Setup

1) System Parameters: We conduct experiments under two different delay settings. The system parameters are set as follows: a three-armed model with *Arm1* being the optimal arm, and the initial preference bias $\theta = [1, 5, 5]$, i.e., the optimal arm has the least initial bias. We choose a three-armed model since large arm set requires a proportional large time horizon to distinguish optimal arm, while in the public Amazon Review Data, the amount of reviews for most products is limited (no more than 3,000 for each product). The self-reinforcing preference function is chosen as $F(x) = x^{\alpha}$ with $\alpha = 2$. The constant incentive for each time step is set as $b = 1.5$ with an incentive impact function $G(b, t) = b$. For the delay distribution, we use normal distributions in both assumption setting, as normal distributions have an infinite support $x \in \mathbb{R}$. Under the arm-independent delay setting, we choose the delay distribution as $\tau_t \sim N(10, 2)$. Under the arm-dependent delay setting, we choose the delay distributions as $\tau_{1,t} \sim N(80, 2)$, $\tau_{2,t} \sim N(10, 2)$, and $\tau_{3,t} \sim N(10, 2)$ for Arms 1, 2, and 3, respectively. We only generate non-negative samples of delay under both assumptions.

Table 4.1: Means of products (arms) in different categories.

Product Category	**Arm1(optimal)**	**Arm2**	**Arm3**
Pet Supplies	0.773	0.656	0.626
Electronics	0.757	0.605	0.617
Home and Kitchen	0.875	0.588	0.673
Books	0.915	0.551	0.706

2) Dataset: We use **Amazon Review Data** (Ni et al., 2019) to provide a practical learning environment. The Amazon Review Data includes 233 million customer reviews (ratings, posting times) for 29 product categories. In the experiment, we

select three products to serve as the arms that have the largest number of reviews in category Pet Supplies, Electronics, Home and Kitchen, and Books, respectively. For each product (arm), we leverage the `rating` and `unixReviewTime` information in each review, and the total number of reviews is 3,000 for each product. The range of ratings in Amazon Review Data is the discrete set $\{1, 2, 3, 4, 5\}$. We convert the rating values to binary by setting the rating values 1 and 2 as 0 and the rating value 4 and 5 as 1, and the reviews with rating value 3 are removed. For each product, the binary review ratings are sorted by `unixReviewTime`, so the ratings come in real-world order in the experiment. We summarize the mean values of the products by their Bernoulli ratings in the four selected categories, as shown in Table 4.1.

4.4.2 UCB-FDF with Arm-Dependent/Independent Delay

The experiment results are illustrated in Figure 4.1-4.4. Figure 4.1 shows the average regret and incentive trends with policy UCB-FDF under setting with no delay. Figure 4.2 and 4.3 show the average regret and incentive trends with policy UCB-FDF under settings with arm-independent delays and arm-dependent delays, respectively. In Figures 4.3 and 4.4, we compare the performances with policy UCB-FDF and baseline policy UCB-List (Zhou et al., 2021). Specifically, Figure 4.4 shows the performance under policy UCB-List in the face of arm-dependent delays that is the same as that in Figure 4.3. Each curve is constructed by regret or incentive values with different time horizons from $T = 150$ to $T = 3000$, incremented by 150. Each node value in curves are averaged by 100 trials.

Discussion: Comparing Figure 4.1 with Figure 4.2 and 4.3, we can observe the delay impact on regret and total incentive, that both the regret and total incentive

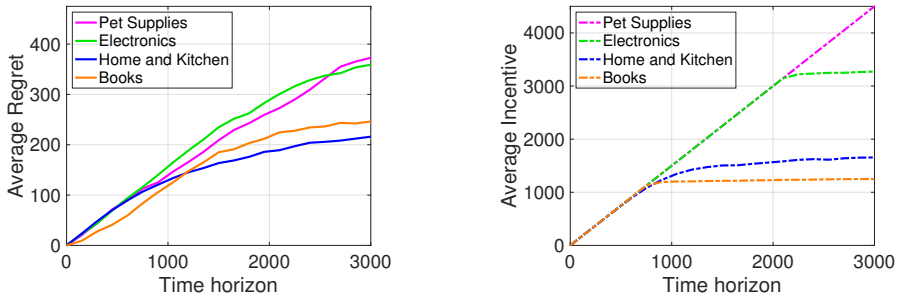


Figure 4.1: The performance of policy UCB-FDF in the face of no delay.

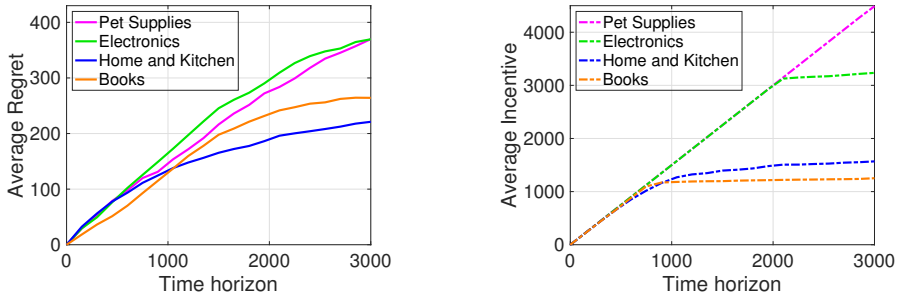


Figure 4.2: The performance of policy UCB-FDF in the face of arm-independent delay.

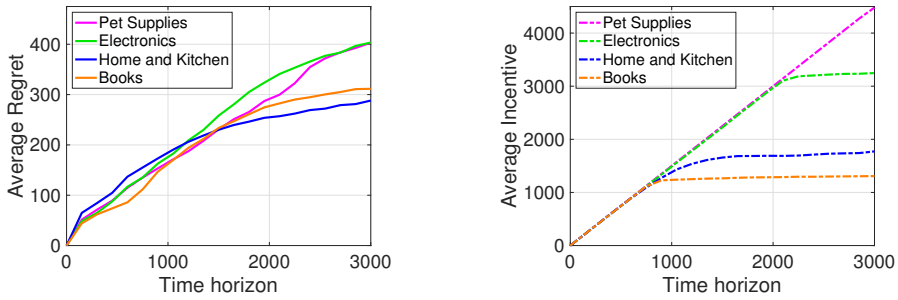


Figure 4.3: The performance of policy UCB-FDF in the face of arm-dependent delay.

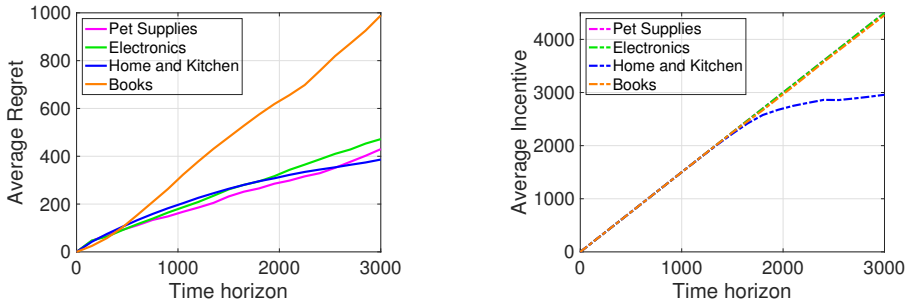


Figure 4.4: The performance of policy UCB-List in the face of arm-dependent delay.

are increased due to the delayed feedback. Comparing Figure 4.3 and Figure 4.4, we observe that under the bandit instances in the face of same delayed feedback, our policy UCB-FDF reaches sub-linear growth rate in both regret and total incentive, except the total incentive in category Pet Supplies, since it may require more time steps to converge while our data is limited, while the policy UCB-List cannot guarantee sub-linear growth rate for both regret and total incentive.

4.4.3 Delay Assumptions Comparisons

In this section, we compare our policy UCB-FDF with UCB-List in Zhou et al. (2021). The system parameters are almost the same as that in Section 4.4.1. We use the three-armed model with the initial preference bias $\theta = [1, 5, 5]$. The self-reinforcing preference function is chosen as $F(x) = x^{\alpha}$ with $\alpha = 2$. The constant incentive for each time step is set as $b = 1.5$ with an incentive impact function $G(b, t) = b$. The experiment results are illustrated in Figure 4.5. We choose the delay distribution as $\tau_{1,t} \sim N(10, 2)$, $\tau_{2,t} \sim N(15, 2)$, and $\tau_{3,t} \sim N(18, 2)$ for arm1, arm2, arm3, respectively. We only generate non-negative samples of delay under both assumptions. (a) and (b) show the average regret and incentive trends with policy UCB-FDF under settings with arm-independent delays and arm-dependent delays, respectively. Each curve is constructed by regret or incentive values with different time horizons from $T = 150$ to $T = 3000$, incremented by 150. Each node value in curves are averaged by 100 trials. Figure (c) illustrates the 100 values of regret or incentive costs with time horizon $T = 3000$ under arm-dependent delays.

Discussions: From the figures, we observe that our UCB-FDF policy achieves sub-linear growth rates for regret and incentive costs under our assumptions of feed-

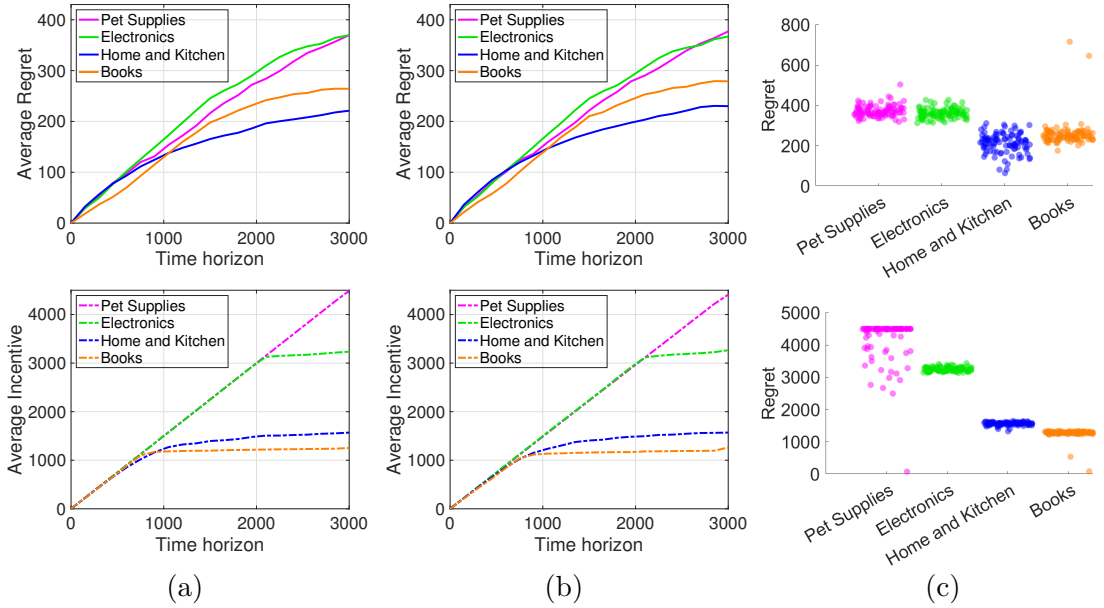


Figure 4.5: Regret and incentive trends of four categories under two different delay settings in (a) and (b). Jittered plot of 100 random cases with $T = 3000$ under arm-dependent delays in (c).

back delays. Comparing results under arm-independent delays in Figure 4.5(a) and arm-dependent delays in Figure 4.5(b), we observe that arm-dependent delays require higher incentive costs and result in larger regret, due to the longer exploration and exploitation phases. Comparing regrets in different categories, we observe that the regret growth trends of categories "Books" and "Home and Kitchen" increase rapidly initially but quickly slow down, since the arms in these two categories have larger gap summations $\sum_{a \neq a^*} \Delta_a$. Although a large gap summation may cause larger regret before the policy finds the optimal arm, it also allows the algorithm to distinguish the optimal arm faster. More straightforward results can be seen by the incentive curves in different categories. From a different angle, as we show in Figure 4.5(c), some of the total incentives in category Pet Supplies are exactly 4500, implying that in many random cases, the policy is still in the exploration or exploitation phase, while in

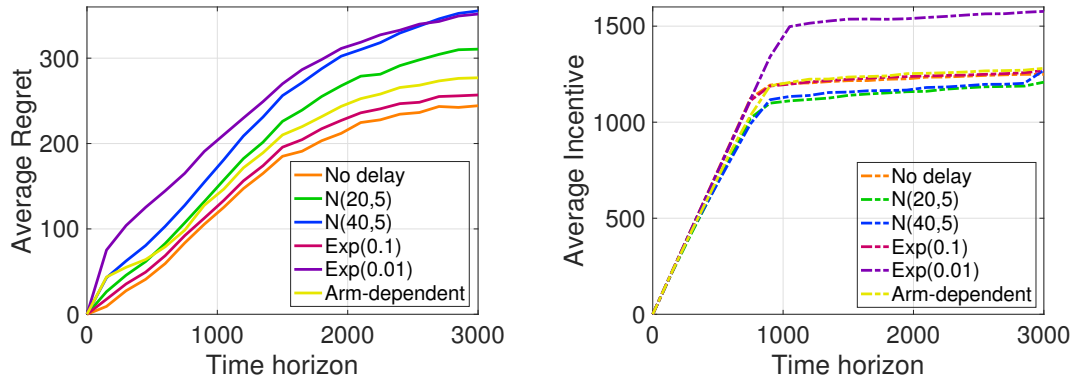
most of the random cases in categories Books and Home and Kitchen, the incentive can stop early while also reach sub-linear regret. This difference is also caused by the variation in gap summations.

4.4.4 Delay Distribution Comparisons

In this section, we compare performance of policy UCB-FDF under different delay distributions using the data in category Books. The experiment results are illustrated in Figure 4.6. (a) in Figure 4.6 shows the different regret under different delay distributions, and (b) shows the corresponding total incentive. For arm-independent delays, we show the policy performance under distribution $N(20, 5)$, $N(40, 5)$, and $Exp(0.1)$, $Exp(0.01)$. For arm-dependent delays, we choose the delay distribution as $\tau_{1,t} \sim N(40, 5)$, $\tau_{2,t} \sim Exp(0.01)$, $\tau_{3,t} = 0$ for arm1, arm2, arm3, respectively.

We also show the average exploration phases ($\bar{t}_1$) and average exploitation phases ($\bar{t}_2$) under four different settings of delay distribution: i) no delay; ii) arm-independent delay with normal distribution $N(40, 5)$; iii) arm-independent delay with exponential distribution $Exp(0.01)$; iv) arm-dependent delay with $\tau_{1,t} \sim N(40, 5)$, $\tau_{2,t} \sim Exp(0.01)$, $\tau_{3,t} = 0$ for arm1, arm2, arm3, respectively. The results are shown in Table 4.2.

Discussion: From Figure 4.6, we observe that the policy performance is impacted by different delay distributions and different distribution parameters. It aligns with our theoretical analysis that the case with no delay has the smallest regret, and as delay expectation getting larger, we have larger expected regret and larger total incentive. From Table 4.2, we firstly have the general observation that as time horizon increases, both the exploration phase and the exploitation phase are increasing, due



(a) (b)

Figure 4.6: Regret of policy UCB-FDF in the face of different delay distributions in (a), and the corresponding total incentive in (b).

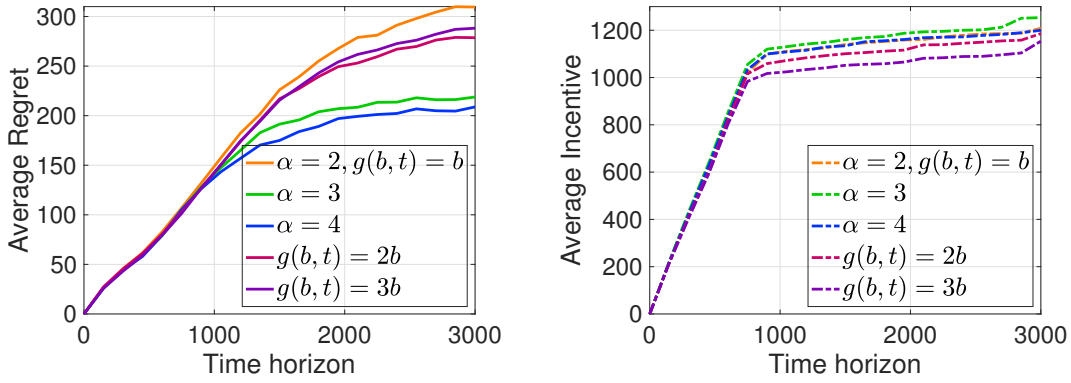
Table 4.2: The average exploration phases ($\bar{t}_1$) and average exploitation phases ($\bar{t}_2$) under four different settings of delay distribution.

Time Horizon		**300**	**600**	**900**	**1200**	**1500**	**1800**	**2100**	**2400**	**2700**	**3000**
$\bar{t}_1$	**No delay**	300	600	781	790	796	803	809	813	816	818
	$N(40,5)$	278	527	691	705	717	724	736	743	748	796
	$Exp(0.01)$	298	595	863	880	896	901	907	915	919	927
	Arm-dependent	279	555	739	760	764	769	780	783	788	800
Time Horizon		**300**	**600**	**900**	**1200**	**1500**	**1800**	**2100**	**2400**	**2700**	**3000**
$\bar{t}_2$	**No delay**	300	600	795	802	811	816	823	826	831	835
	$N(40,5)$	291	539	745	759	771	777	787	793	798	847
	$Exp(0.01)$	299	596	894	1010	1024	1024	1030	1038	1043	1051
	Arm-dependent	290	567	793	816	823	827	836	841	846	854

to a more accurate estimation over arms. On the other hand, we observe that the length of exploration phase is problem-dependent, while the length of exploitation phase is positive proportional with the value of expected delay, which is due to our exploitation mechanism where dominance threshold grows with expected delay.

4.4.5 Comparison with Different Parameters

In this section, we compare the performance of policy UCB-FDF under different system parameters using the data in category Books, specifically, we compare the impact of different strengths of feedback function $F(\cdot)$ and incentive impact function $G(\cdot)$. We use the setting $F(x) = x^{\alpha}$ with $\alpha = 2$ and $G(b,t) = b$ as the comparison group. For the other four groups of experiment, we use $\alpha = 3$ and $\alpha = 4$ for feedback function with $G(b,t) = b$, and $G(b,t) = 2b$ and $G(b,t) = 3b$ for incentive impact function with $\alpha = 2$, respectively. (a) in Figure 4.7 shows the different regret under different parameters, and (b) shows the corresponding total incentive.



(a) (b)

Figure 4.7: Regret of policy UCB-FDF in the face of different feedback functions and incentive impact functions in (a), and the corresponding total incentive in (b).

Discussion: For the feedback function $F(\cdot)$, as α increases, the strength of self-reinforcing preferences is enhanced, which implies that the preferences are easier to be induces to one arm, so that we observe smaller regret. For the incentive impact function $G(b,t) = c \cdot b$, as the coefficient c increases, the preference impact of a unit payment is enhances, implying that users are easier to be affected (controlled) by the

incentive, thus we observe a slightly smaller regret, with obviously decreased total incentive.

Chapter 5

Bandit Learning to Rank with Position-Based Click Models: Personalized and Equal Treatments

5.1 Overview

A key component in MAB-based ONL2R is the click model. A natural choice of click model is the so-called *position-based click model* (Richardson et al., 2007; Lagrée et al., 2016), where each ranking position is associated with a preference probability of being observed and clicked. Studies have shown that user actions are highly influenced by webpage layouts or ranking positions: if a listing is not displayed in some particular area of the web layout, then the odds of being seen by a searcher are dramatically reduced (Hotchkiss et al., 2005). Position-based click model is also shown to be closely related to various popular ranking quality metrics for recommendation systems, such as normalized discounted cumulative gain (NDCG)(Valizadegan et al., 2009; Wang et al., 2013).

However, developing efficient online learning policies for MAB-based ONL2R with position-based click models is highly non-trivial due to the following technical challenges: First, MAB-based ONL2R problems with position-based click models are *combinatorial* in nature, which means that their offline counterparts are already NP-hard in general. Second, due to the multi-user nature, ranking recommendations for MAB-based ONL2R problems are complicated by the philosophical debate whether

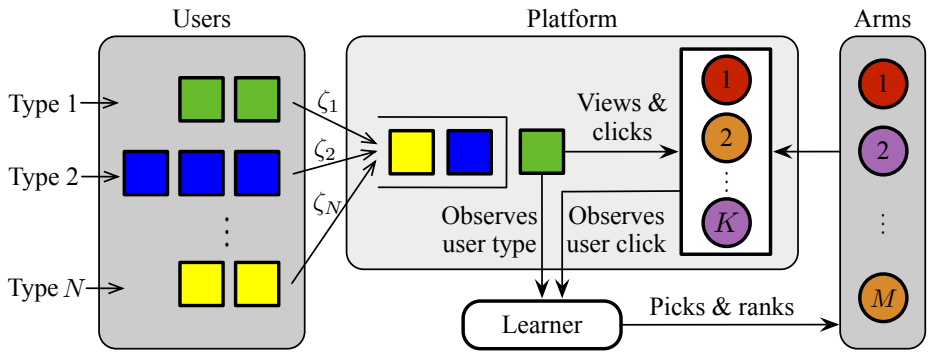


Figure 5.1: System model of MAB-based ONL2R with position-based click models.

we should provide personalized or equal treatments to different users, both of which are common in practice. To date, there remains a lack of rigorous understanding on how different types of ranking treatments could affect MAB-based ONL2R policy design. Third, unlike conventional MAB problems, there is a fundamental *partial observability* challenge in MAB-based ONL2R policy design. Specifically, in many real-world recommendation systems, if a user does not click on any displayed ranked item, the system will not receive any feedback on which ranked position has been observed by the user. This uncertainty creates an extra layer of challenge in MAB-based ONL2R. Last but not least, in MAB-based ONL2R with position-based click models, there is a complex coupling between each ranking position's observation preference and mean reward of each arm, both of which are not only unknown and need to be learned, but also heterogeneous across user types. Due to these challenges, results for MAB-based ONL2R with position-based click models are rather limited in the literature, which motivates us to fill this gap.

5.2 System Model and Problem Formulation

1) System Setup: As shown in Fig. 5.1, consider a stochastic bandit setting with a set of user types $[N] \coloneqq \{1, \ldots, N\}$, a set of arms $[M] \coloneqq \{1, \ldots, M\}$, and a set of ranking positions $[K] \coloneqq \{1, \ldots, K\}$, where $K \leq M$. Each user type i has an arrival rate $\zeta_i > 0$. Without loss of generality, the arrival rates are normalized such that $\sum_{i \in [N]} \zeta_i = 1$. For each position $k \in [K]$, each user type $i \in [N]$ has a position preference $\rho_{i,k} \in [0, 1]$, which represents the chance that user type i observes position k. We note that such position preferences have been widely observed in practice. For example, demographic studies Hotchkiss et al. (2005) showed that there exist many different user's position-based action patterns that are related to user's gender, education, age, etc. Such user-specific position-based patterns include "quick click," "the linear scan," "the deliberate scan," "the pick up search," etc. Thus, the same position can have different preference rates over different groups of people. For each arm $j \in [M]$, each user type i has a Bernoulli reward distribution $D_{i,j}$ with mean $\mu_{i,j} \in [0, 1]$, where $\mu_{i,j}$ can be interpreted as the click rate of arm j if observed by user type i. We assume that the arrival rates ζ_i, the position preferences $\rho_{i,k}$, and the arm means $\mu_{i,j}$ are all *unknown* to the learner. With the basic system setup, we are now in a position to describe the unique key features of our MAB model.

2) Agent-User Interaction Protocol: At time step t, a user of type $I(t) = i$ arrives with probability λ_i. With a slight abuse of notation, we also use $I(t)$ to denote the current user at time step t. The learning agent observes $I(t)$, then picks a K-sized subset of arms from $[M]$ and determines a K-permutation $\sigma_t \in P_K^M$, where $\sigma_t(j)$ represents the ranked position of arm j. Next, user $I(t)$ randomly observes an arm $J(t)$ at ranked position $\sigma_t(J(t))$ with probability $\rho_{I(t),\sigma_t(J(t))}$, and then clicks arm $J(t)$

with probability $\mu_{I(t),J(t)}$. The learner receives a reward $X(t) = 1$ if some position is clicked, and $X(t) = 0$ otherwise. Clearly, we have $\sum_{k\in[K]} \rho_{i,k} = 1$. We note that if user $I(t)$ chooses not to click any arm $J(t)$, then the learner receives *no information* regarding which position has been observed. In other words, a reward of $X(t) = 0$ can happen by the fact that any random arm in σ_t is observed but not clicked. This partial observation setting closely follows the reality in most recommendation systems, while it also makes the estimation of arm means and position preferences much harder. We illustrate this learner-user interaction in Fig. 5.1. A policy π sequentially makes decisions on the permutation σ_t, and observes stochastic rewards $X(t)$ over time t.

3) Regret Modeling: In this work, we consider two MAB-based ONL2R problems with two different ranking recommendation settings: personalized and equal treatments.

3-a) Personalized Treatment: In this setting, the learning policy makes ranking personalized decisions according to the arrived user types. Thus, an optimal policy always recommends the optimal permutation denoted by σ_i^* regarding the arrived user type i to achieve *maximum user satisfaction*. We note, however, that personalized treatment may not be fair to the arms since some arms may never be shown to any user type. At time t, an expected regret $\mathbb{E}[R(t)]$ incurred by a policy $\{\sigma_t\}_t$ is defined as follows:

$$\mathbb{E}[R(t)] = \sum_{s=1}^{t} \left(\langle \boldsymbol{\rho}_{I(s)}, \boldsymbol{\mu}_{I(s),\sigma^*} \rangle - \mathbb{E}[\langle \boldsymbol{\rho}_{I(s)}, \boldsymbol{\mu}_{I(s),\sigma_s} \rangle] \right),$$

where we use the notation of vector $\boldsymbol{\rho}_i = [\rho_{i,1}, \rho_{i,2}, \ldots, \rho_{i,K}]^\top$, and use the notation $\boldsymbol{\mu}_{i,\sigma} = [\mu_{i,\sigma^{-1}(1)}, \ldots, \mu_{i,\sigma^{-1}(K)}]^\top$ to denote the vector of arm means ranked by permutation σ, where $\sigma^{-1} : [K] \to [M]$ is the reverse mapping of the permutation σ.

3-b) Equal treatment: Motivated by fairness, an equal treatment ranking policy makes an identical ranking decision among all user groups to avoid discrimination over sensitive groups. This is because, in some scenarios, service providers are legally required to treat all sensitive groups the same way by recommending the same ranking content. In the equal treatment setting, we measure its ranking quality by a general collective utility function (CUF) defined as follows:

$$\Gamma(\sigma) \triangleq \sum_{i\in[N]} \zeta_i \cdot U\bigg(\sum_{j\in \mathrm{M}_\sigma} \rho_{i,\sigma(j)} \cdot \mu_{i,j}\bigg),$$

where M_σ is the set of arm indices in permutation σ, and $U(\cdot)$ is a generic utility function that transforms the CUF to different social welfare criteria, which we will discuss later. An optimal policy always recommends a universal optimal permutation denoted by σ^* that maximizes CUF, i.e., $\sigma^* = \arg\max_{\sigma\in P_K^M} \Gamma(\sigma)$. We note that under this setting, due to multiple user types and their unequal arm means, the optimal permutation σ^* may not be a decreasingly ordered arm list. At time t, an expected regret $\mathbb{E}[R(t)]$ incurred by a policy $\{\sigma_t\}_t$ is defined as follows:

$$\mathbb{E}[R(t)] = t \cdot \Gamma(\sigma^*) - \sum_{s=1}^{t} \mathbb{E}\left[\Gamma(\sigma_s)\right].$$

Here, we provide two common examples of CUF: i) the utilitarian CUF and ii) the Nash CUF (Ramezani and Endriss, 2009). Specifically, let v_i denote the individual user utility. Then, the utilitarian CUF is defined as $\sum_i v_i$, which favors users with higher average utility. The Nash CUF is defined as $\sum_i \log(v_i)$, which balances efficiency and fairness.

5.3 Policy Designs and Performance Analysis

In this section, we focus on policy design and analysis for MAB-based ONL2R with both personalized and equal treatment settings. While these two settings are different, they share some common subtasks (e.g., estimations of position preferences and arm means). Thus, we will consider these common subtasks as preliminaries in Section 5.3.1 first, which paves the way for presenting our policies in Section 5.3.2. Lastly, we will conduct regret analysis for our proposed policies in Section 5.3.3.

5.3.1 Preliminaries

1) Notations and Terminologies: Before we present our proposed bandit policy designs, we introduce some notations and terminologies as follows. At time t, if the policy picks a permutation σ_t, then any arm $j \in \sigma_t$ is said to have been "pulled" by the agent at time t. We use $T_{i,j,k}(t)$ to denote the cumulative pulling times of arm $j \in [M]$ that is offered to users of type $i \in [N]$ at position $k \in [K]$ up to time t, i.e., $T_{i,j,\sigma_t(j)}(t) = T_{i,j,\sigma_t(j)}(t-1) + 1$ for $i = I(t), j \in \mathrm{M}_{\sigma_t}$. Likewise, we use $S_{i,j,k}(t)$ to denote the cumulative reward of arm $j \in [M]$ that is clicked by users of type $i \in [N]$ at position $k \in [K]$, i.e., $S_{i,j,\sigma_t(j)}(t) = S_{i,j,\sigma_t(j)}(t-1) + r(t)$ for $i = I(t), j = J(t)$. We represent tensors using bold notation, e.g., $\boldsymbol{S}_{i,j}(t) = (S_{i,j,1}(t), \ldots, S_{i,j,K}(t))$. We use the notation $\|\cdot\|_1$ to represent the ℓ^1-norm of tensors, e.g., $\|\boldsymbol{S}_{i,j}(t)\|_1 = \sum_{k=1}^{K} S_{i,j,k}(t)$.

The main statistical challenge in our model is to estimate the position preferences and arm means only by the user feedback, i.e., a sequence of the joint realizations of position preference distribution and arm reward distribution. To this end, we present two estimators that detangle the joint realization and estimate the unkown parameters with asymptotic confidence over time.

Algorithm 4 The Position Preference Estimator $E(\boldsymbol{T}(t), \boldsymbol{S}(t))$.

1: **Input:** $\boldsymbol{T}(t), \boldsymbol{S}(t)$

2: **for** all player i, arm j, and position k **do**

3: $\quad \bar{v}_{i,j,k}(t) = \dfrac{S_{i,j,k}(t)/T_{i,j,k}(t)}{\sum_{l\in[K]} S_{i,j,l}(t)/T_{i,j,l}(t)}$

4: **end for**

5: **for** all player i and position k **do**

6: $\quad \hat{\rho}_{i,k}(t) = \dfrac{1}{M}\sum_{j\in[M]} \bar{v}_{i,j,k}(t)$

7: **end for**

2) Position Preference Estimator: The position preference is estimated based on the fact that given a user type i and an arm j, the value $S_{i,j,k}(t)/T_{i,j,k}(t)$ at any position k is asymptotically approaching its expectation $\mu_{i,j} \cdot \rho_{i,k}$ over time. Thus, intuitively, its normalization over all positions asymptotically removes the impact of arm mean on the position preference estimation, as stated in Algorithm 4. It is then possible to obtain a concentration on the estimated position preference, as stated in Lemma 10 below. Due to space limitation, the proofs of all theoretical results are relegated to the supplemental material.

Lemma 10. For each user type $i \in [N]$ and position $k \in [K]$, for any constant $\epsilon \geq 0$, the position preference estimator $E(\boldsymbol{T}(t), \boldsymbol{S}(t))$ achieves a concentration bound as follows:

$$\mathbb{P}\left(|\hat{\rho}_{i,k}(t) - \rho_{i,k}| \geq \max_{j\in[M]} \sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \leq MKt^{-2\epsilon}. \tag{5.1}$$

3) Arm Mean Estimator: For user type i and arm j, we note that directly estimating the arm mean $\mu_{i,j}$ by the value $\|\boldsymbol{S}_{i,j}(t)\|_1/\|\boldsymbol{T}_{i,j}(t)\|_1$ could be biased since

its expectation is different from the arm mean $\mu_{i,j}$. To address this problem, we define an asymptotically unbiased total pulling time estimator $N_{i,j}(t) = \sum_{k\in[K]} T_{i,j,k}(t) \cdot \hat{\rho}_{i,k}(t)$ for $i \in [N], j \in [M]$, which has an increment of $\hat{\rho}_{i,k}(t)$ once pulled at time t and can be leveraged to estimate arm means only with the knowledge of joint realizations. We note here that $N_{i,j}(t)$ is a random variable that depends on both the ranking history $\{\boldsymbol{\sigma}(t)\}_t$ and the position preference distribution. Then, for user type i and arm j, the asymptotically unbiased arm mean estimator can be defined as $\hat{\mu}_{i,j}(t) = \|\boldsymbol{S}_{i,j}(t-1)\|_1 / N_{i,j}(t-1)$.

Following Lemma 10, define event $\mathcal{N}_t$ as follows: at time t, for user i and arm j, there exists $\epsilon \geq 0$ such that $|N_{i,j}(t) - \bar{N}_{i,j}(t)| < \|\boldsymbol{T}_{i,j}(t)\|_1 \max_{j\in[M]} \sqrt{\epsilon \ln t / \left(\mu_{i,j}^2 T_{i,j,k}(t)\right)}$. Then, we have:

Lemma 11. For user type i and arm j, denote the unbiased empirical arm means by $\hat{\mu}_{ij}(t) = \|\boldsymbol{S}_{i,j}(t-1)\|_1 / N_{ij}(t-1)$. Then, conditioned on event $\mathcal{N}_t$ and for any $\epsilon \geq 0$ we have

$$\mathbb{P}\left(\left| \hat{\mu}_{i,j}(t) - \mu_{ij} \right| \geq \frac{\sqrt{2\epsilon \|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{i,j}(t)} \;\middle|\; \mathcal{N}_t \right) \leq \epsilon e^{1-\epsilon} \log t. \tag{5.2}$$

4) CUF Estimator: Different from personalized treatment, equal treatment policies recommend identical ranking lists to all user types, which require extra estimation of user arrival rates. Thus, obtaining an optimal arm ranking order is typically more complicated than obtaining a decreasingly ordered arm list as in personalized treatment. We measure the quality of a permutation by a CUF estimator. Combining $\hat{\rho}_{i,k}(t)$ and $\hat{\mu}_{i,j}(t)$ with the fact that we can estimate the arrival rate of user type i at time t as $\hat{\zeta}_i(t)/\|\hat{\boldsymbol{\zeta}}(t)\|_1$, where $\hat{\zeta}_i(t)$ is the cumulative number of arrived users in type i up to time t, we can estimate all unknown parameters $\rho_{i,k}$, $\mu_{i,j}$, and ζ_i with

asymptotic confidence. Given a ranking σ at time t, we can estimate the unknown CUF $\Gamma(\sigma)$ as follows:

$$\hat{\Gamma}_t(\sigma) = \sum_{i\in[N]} \frac{\hat{\zeta}_i(t)}{\|\hat{\boldsymbol{\zeta}}(t)\|_1} \cdot U\Big(\sum_{j\in \mathrm{M}_\sigma} \hat{\rho}_{i,\sigma}(j)\cdot\hat{\mu}_{i,j}(t)\Big). \tag{5.3}$$

5.3.2 Policy Design for Personalized and Equal Treatments

Based on the estimators presented in Section 5.3.1, we are now in a position to present our policy designs for both personalized and equal treatments.

1) GreedyRank: GreedyRank is a greedy policy that has an increasing probability to exploit the empirical best permutation over time and a decreasing probability to explore combinations of every arm and position in a round-robin fashion. The policy is described in Algorithm 5.

Algorithm 5 The GreedyRank Policy.

Input: ε_t
Initialization: $\sigma_t = \big\{\text{arm} \to \text{position} : [(t+k) \bmod M] + 1 \to k, k \in [K]\big\}$ till $\min_{i,j,k} S_{i,j,k}(t) > 0$
Mark current time $t_0 = t$, and let $I_{\text{explore}} \leftarrow 1$
for $t = t_0 + 1, \ldots$ **do**
    Observe user type $I(t)$, and toss a coin with head rate of ε_t
    if head **then**
        $\sigma_t = \big\{\text{arm} \to \text{position} : [(I_{explore} + k) \bmod M] + 1 \to k, k \in [K]\big\}$
        $I_{\text{explore}} = (I_{\text{explore}} \bmod M) + 1$
    else
        Option 1 [Personalized treatment]: $\sigma_\mu \leftarrow$ decreasingly rank arm indices by $\hat{\mu}_{I(t),j}(t)$
                $\sigma_\rho \leftarrow$ decreasingly rank position indices by $\hat{\rho}_{I(t),k}(t)$
                $\sigma_t = \{\text{arm} \to \text{position} : \sigma_\mu(a) \to \sigma_\rho(a), a \in [K]\}$
        Option 2 [Equal treatment]: $\sigma_t = \arg\max_{\sigma\in P_K^M} \hat{\Gamma}_t(\sigma)$
    end if
    Observe user feedback $X(t)$, update parameters $S_{i,j,k}(t)$, $T_{i,j,k}(t)$, $\hat{\boldsymbol{\rho}}(t)$, $N_{i,j}(t)$ as stated in Section 5.3.1
end for

2) UCBRank: Under UCBRank, the personalized treatment allows UCB-style policies to sort optimistic indices in a decreasing order and pick the corresponding first K arms as permutation, while the equal treatment searches for a the permutation that maximizes the estimated CUF and a confidence interval. To balance exploration and exploitation, we use a confidence interval derived from the McDiarmid's inequality, as described in Algorithm 6.

Algorithm 6 The UCBRank Policy.

Input: a_t
Initialization: $\sigma_t = \{\text{arm} \rightarrow \text{position} : [(t+k) \bmod M] + 1 \rightarrow k, k \in [K]\}$ till $\min_{i,j,k} S_{i,j,k}(t) > 0$
Mark current time $t_0 = t$
for $t = t_0 + 1, \ldots$ **do**
    Observe user type $I(t)$
    Option 1 [Personalized treatment]: $\sigma_\mu \leftarrow$ decreasingly rank arm indices by $\hat{\mu}_{I(t),j}(t) + \dfrac{a_t \ln t}{N_{I(t),j}(t)}$
        $\sigma_\rho \leftarrow$ decreasingly rank position indices by $\hat{\rho}_{I(t),k}(t)$
        $\sigma_t = \{\text{arm} \rightarrow \text{position} : \sigma_\mu(a) \rightarrow \sigma_\rho(a), a \in [K]\}$
    Option 2 [Equal treatment]: $\sigma_t = \arg\max_{\sigma \in P_K^M} \left(\hat{\Gamma}_t(\sigma) + \sum_{i \in [N]} \sum_{j \in \mathrm{M}_\sigma} \dfrac{a_t \ln t}{N_{i,j}(t)} \right)$
    Observe user feedback $X(t)$, update parameters $S_{i,j,k}(t)$, $T_{i,j,k}(t)$, $\hat{\boldsymbol{\rho}}(t)$, $N_{i,j}(t)$ as stated in Section 5.3.1
end for

One important remark regarding the time complexity of the equal treatment option in both policies is in order. In both policies, The option of equal treatment requires solving an integer (combinatorial) optimization problem, which may be NP-hard depending on the type of the utility function U. As a result, the equal treatment setting is more challenging in general than the personalized treatment setting in MAB-based ONL2R. Also, it is easy to see that the search space of the optimization is $\mathcal{O}(M^K)$. To this end, several interesting cases may happen:

1) *K is Fixed and Moderate:* This case corresponds to a short ranking list (e.g., due to limited screen space on phones). In this case, the search space is polynomial with respect to M. Thus, the integer optimization problems in Lines 11 and 7 in Algorithms 5 and 6, respectively (referred to as IOPs) can be exactly solved by brute force search with acceptable time complexity. However, if M is fixed or K is fixed but large, a brute force search is clearly too costly.

2) *U is Linear and the M^K-value Is Moderate:* This case happens if utilitarian criterion is used. In this case, IOPs can be equivalently transformed to an integer linear program (ILP) over a probability simplex (associating each permutation with a binary variable such that these binary variables sum to 1). Thanks to this simple structure, solving the LP relaxation of the transformed problem automatically yields a binary solution (hence the optimal ranking) in polynomial time.

3) *U is Linear and the M^K-value Is Large (Exponential):* In this case, even solving the transformed LP relaxation of the optimization problems (Lines 11 and 7 in Algorithms 5 and 6, respectively) could still be cumbersome due to the large problem size. One solution approach is to uniformly sample with probability p a subset of all possible permutations and then solve an LP relaxation only using the sampled permutations. Clearly, as $p \to 1$, the LP solution will be arbitrarily close to the optimal solution, hence yielding a *polynomial-time approximation scheme* (PTAS).

4) *U is Concave:* This case happens if, e.g., Nash criterion ($U = \log(\cdot)$) is used. In this case, IOPs are an integer convex optimization problem, which may be solved by state-of-the-art branch-and-bound-type (BB) optimization scheme Sierksma and Zwols (2015) to high accuracy relatively fast, if exponentially growing running

time is tolerable.

5) *U is Non-Concave:* In this case, IOPs are an integer non-convex problem (INCP), which is the hardest. Although one can still use branch-and-bound-type schemes in theory, the time complexity could be arbitrarily bad. However, developing INCP algorithms is beyond the scope of this paper.

5.3.3 Regret Analysis

We now present the theoretical results on the proposed policies, as well as their proof sketches. The complete proofs are provided in the supplemental material.

Theorem 12. *(Personalized treatment with GreedyRank)* Setting $\varepsilon_t = t^{-1/2}$, the expected regret of GreedyRank Option 1 at any time step t can be bounded as follows:

$$\mathbb{E}[R(t)] \leq 2N\sqrt{t} + \sum_{i\in[N]} \frac{8C_\rho MK\sqrt{\zeta_i t}\ln t}{(1-1/C)\min_j \mu_{i,j}} + \mathcal{O}(1),$$

where $C, C_\rho > 1$ are problem-dependent constants.

Proof Sketch of Theorem 12. Based on the lemmas presented in Section 5.3.1, we define two "good" events $\mathcal{P}_t$ and $\mathcal{U}_t$ regarding the estimated position preference and the estimated arm mean, respectively. Then, for user type i and arm j, conditioned on events $\mathcal{P}_t$ and $\mathcal{U}_t$, we obtain a concentration bound regarding the quantity $\hat{\mu}_{i,j}(t) \cdot \hat{\rho}_{i,\sigma_t(j)}(t)$ for any policy σ_t. Next, we define another "good" event $\mathcal{F}_t$ regarding the minimum cumulative number of exploration times that GreedyRank performs up to time t. Conditioned on all the defined events, we obtain the upper bound of the regret $\mathbb{E}[R(t)]$ with respect to ε_t, and the upper bound of $\mathbb{E}[R(t)]$ is minimized with

$\varepsilon_t = t^{-1/2}$. Finally, upper bounding the complementary of all the conditioned events finishes the proof. □

Theorem 13. *(Personalized treatment with UCBRank)* Setting $a_t \in (2/\min \mu_{i,j}, \sqrt{t/\ln t}]$, the expected regret of UCBRank Option 1 at any time step t can be bounded as follows:

$$\mathbb{E}[R(t)] \leq \sum_{i \in [N]} \frac{2C_\rho MK\sqrt{\zeta_i t \ln t}}{\min_j \Delta_{i,j}} + \mathcal{O}(1).$$

Proof Sketch of Theorem 13. We define a "bad" event $\mathcal{E}_t$ regarding the appearance of sub-optimal permutation selection. Then, we show that event $\mathcal{E}_t$ happens with probability zero if i) events $\mathcal{P}_t$ and $\mathcal{U}_t$ happens, ii) a_t is in a proper range, and iii) each possible permutation has been sampled for enough times. By bounding the probability of the appearance of event $\mathcal{E}_t$ for each time, and bounding the complementary of all the conditioned events as in the proof of Theorem 12, we finish the proof. □

For the equal treatment option in both policies, to address the potential NP-Hardness challenge in solving the optimization problems (Lines 11 and 7 in Algorithms 5 and 6, respectively), we consider approximation algorithms to balance the trade-off between time complexity and optimality. Also, compared to personalized treatment, equal treatment policies are more expensive on average due to the extra estimation of user arrival rate.

To avoid linear regret, we require the continuity assumption on the utility function f as follows:

Assumption 2 (Bi-Lipschitz Continuity)**.** The function U is L_U bi-Lipschitz continuous, i.e., there exists a constant $L_U \geq 0$ such that for any $x_1, x_2 \in [\min_{i,j,k}\{\rho_{i,k}\mu_{i,j}\}, \max_{i,j,k}\{\rho_{i,k}\mu_{i,j}\}]$, it holds that $(x_1 - x_2)/L_U \leq U(x_1) - U(x_2) \leq L_U(x_1 - x_2)$.

Theorem 14. *(Equal Treatment with GreedyRank)* Setting $\varepsilon_t = Nt^{-1/2}$, with a δ_t-approximate solution to the maximization problem in GreedyRank Option 2, the expected regret of Fair-GreedyRank at any time step t can be bounded by:

$$\mathbb{E}[R(t)] \leq 2Nt^{\frac{1}{2}} + \frac{8L_U C_\rho NMK}{(1-1/C)\min \mu_{i,j}} t^{\frac{1}{2}} \ln t + \sum_{s=1}^{t} U(1) N \delta_s + \mathcal{O}(1),$$

and for $\delta_t = \mathcal{O}(t^{-1})$, we have: $\mathbb{E}[R(t)] = \mathcal{O}\left(8L_U NMKt^{\frac{1}{2}} \log t / \min \mu_{i,j}\right)$.

Proof Sketch of Theorem 14. In the equal treatment setting, we obtain a concentration on the estimated CUF $\hat{\Gamma}_t(\sigma)$ by i) the properties of the utility function U, and ii) the estimated user arrival rate $\hat{\zeta}_i(t)$. Note that $\hat{\zeta}_i(t)$ can be bounded by Hoeffding's inequality. Then, when bounding the regret upper bound in a similar manner as that in the proof of Theorem 12, we additionally consider the suboptimality from the approximated solution, which causes an extra regret of at most $U(1)N\delta_t$ for each time t. □

To present the regret result of equal treatment UCBRank, we define the minimum reward gap Δ_Γ as follows: $\Delta_\Gamma = \Gamma(\boldsymbol{\sigma}^*) - \max_{\boldsymbol{\sigma} \in P_K^M, \boldsymbol{\sigma} \neq \boldsymbol{\sigma}^*} \Gamma(\boldsymbol{\sigma})$. We note that Δ_Γ only depends on the distributions of user arrival rate $\boldsymbol{\zeta}$, position preference $\boldsymbol{\rho}$, and arm mean $\boldsymbol{\mu}$.

Theorem 15. *(Equal Treatment with UCBRank)* With any δ_t-approximate solution to the maximization problem in UCBRank Option 2, setting $\delta = \mathcal{O}(\sqrt{\log t / t})$ and $a_t \in (2L_U / \min \mu_{i,j}, \sqrt{t / \ln t}]$, the expected regret of UCBRank at any time step t can

be bounded as follows:

$$\mathbb{E}[R(t)] = \mathcal{O}\left(\frac{N^2MK\sqrt{t\log t}}{\Delta_\Gamma}\right).$$

Proof Sketch of Theorem 15. Similar as the proof of Theorem 13, we define a "bad" event $\mathcal{E}_t$ regarding the appearance of sub-optimal permutation selection, and we show that event $\mathcal{E}_t$ happens w.p.0 if two additional conditions are satisfied: i) the estimated CUF $\hat{\Gamma}_t(\sigma_t)$ is accurate enough, which is shown in the proof of Theorem 14, and ii) the suboptimality of the approximation δ_t is upper bounded. □

5.4 Experiments

5.4.1 Experiment on Synthetic Data

We first conduct experiment on a synthetic dataset, where $N = 3$, $M = 20$, and $K = 4$. We relax the assumption of Bernoulli distributed arm means, and replace it by the Beta distribution, which allows discrete-valued reward with a scaled-up arm expectation (this will increase the regret while it helps the estimation with a large problem size in a finite-time horizon). All the system parameters are randomly sampled. We set $\varepsilon_t = a_t = 1$ for personalized treatment GreedyRank (PT-GreedyRank) and UCBRank (PT-UCBRank), respectively. We set $\varepsilon_t = a_t = 5$ for equal treatment GreedyRank (ET-GreedyRank) and UCBRank (ET-UCBRank), respectively. Since there is no existing algorithm that works in our context, we set a baseline with a idea similar to most existing algorithms: the baseline runs the UCB algorithm that shares the same confidence interval as ours in PT-UCBRank, while the baseline does

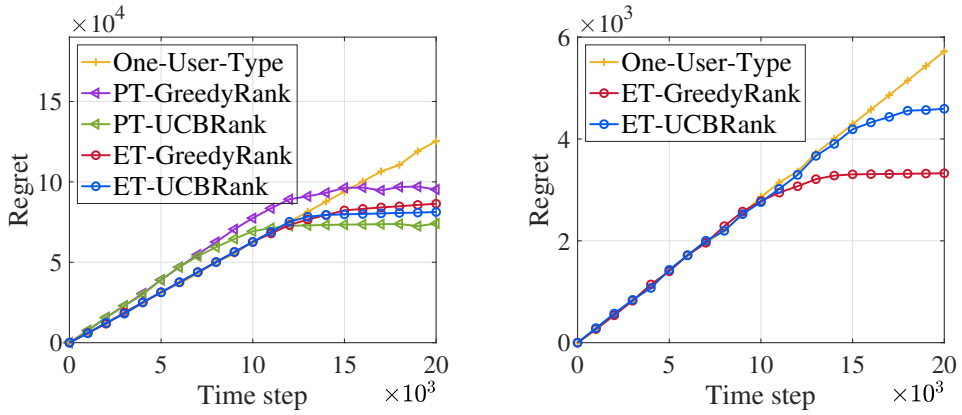


Figure 5.2: Baselines, personalized policies (left), equal treatment policies in utilitarian CUF (left) and Nash CUF (right) on synthetic dataset.

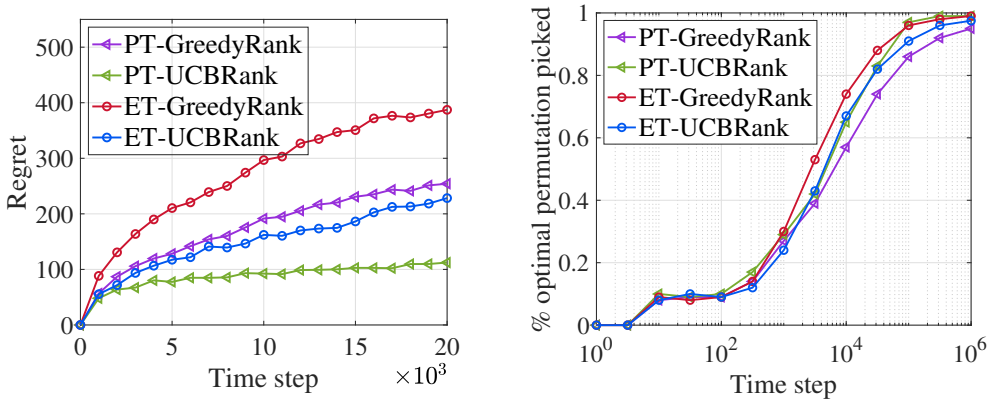


Figure 5.3: Average regret of proposed policies (left), and the optimal action rate of proposed policies (right) on real-world dataset.

not distinguish different user types and treats all users as one type. We present the results in Figures 5.2 and 5.3. The curves confirm our analysis that all proposed policies are sub-linear in regret, and the performance of each policy also depends on the system parameters and policy parameters, e.g., ET-GreedyRank is optimal in utilitarian CUF but not in Nash CUF.

5.4.2 Experiment on Real-World Data

We use the dataset provided for KDD Cup 2012 track 2 [1], which is about advertisements shown alongside search results in a search engine owned by Tencent. The users in the dataset are numbered in millions, and are provided with demographics information, e.g., gender. Ads are displayed in a position (1, 2, or 3) with a binary reward (click or not). Since ads are rarely displayed in position 3, which results in a lack of data, so we focus on two positions (1 and 2). We pick the top 5 ads with high frequency and present the statistical information in Table A.3. We set $\varepsilon_t = a_t = .25$ for PT-GreedyRank and PT-UCBRank, respectively. We set $\varepsilon_t = a_t = .5$ for ET-GreedyRank and ET-UCBRank, respectively. We present the results in Fig. 5.3. The results show that all the proposed policies find the optimal permutations over time. Although the rewards are synthetic, this experiment is still realistic since the values of all other parameters are extracted from the real world.

Table 5.1: Statistics on the dataset, including two user types (male and female), user arrival rate, arm means and position preference (bias).

	Arm Mean					Position Bias	
Gender	1	2	3	4	5	1	2
Male	.357	.471	.604	.808	.564	.323	.677
Female	.247	.327	.491	.49	.303	.416	.584
Arrival Rate: .52(M) : .48(F)							

We also compare the case where approximation algorithms are used in equal treatment ranking. we use a PTAS for utilitarian CUF maximization in both policies: given a ratio δ_t that gets close to one over time, randomly sample $\delta_t n(P_K^M)$ permutations and find the optimal in the samples. If the utility function U is linear, it can be

[1] https://www.kaggle.com/datasets/mohamedkhaledelsafty/click-prediction

Table 5.2: Comparison of equal treatment policies with approximated solution and optimal solution under utilitarian CUF.

		$t = 3 \cdot 10^5$		$t = 6 \cdot 10^5$	
		regret	running time/s	regret	running time/s
GreedyRank	approx.	512	33	581	68
	opt.	387	41	461	87
UCBRank	approx.	314	75	326	162
	opt.	238	82	249	188

easily shown that this strategy is PTAS. The experiment on real-world dataset runs on CPU configured by Apple M1 with 8-core and 3.2 GHz, with 16 GB main memory. The results in Table A.2 confirm our analysis on the tradeoff between regret and time complexity.

Chapter 6

Discussions and Conclusion

6.1 Summary

Multi-armed bandit is a critical model in online learning. As research in recommendation systems attracts more attention in academia, it is necessary to understand and analyze user behavior formulated in bandit framework, so as to design efficient policies that achieve long-term objectives of service providers. To this end, in this thesis, we studied bandit models that are applicable in real-world recommendation systems. We tackled in the following two aspects:

(1) We modeled commonly seen user behavior in recommendation systems by making natural assumptions and reasonable formulations, and we modeled arm filtering strategies that are well-adopted by service providers of recommendation systems.

(2) We designed efficient policies that aim to maximize long-term reward of service providers, and theoretically show that all the proposed policies achieve sublinear regret.

Specifically, in (1), we formulated self-reinforcing user preference and theoretically analyzed how user preference evolves by urn model, and proved an interesting property of arm dominance, which is the critical property for subsequent policy design. Then,

we analyzed the impact of delayed user feedback on the basis of model with self-reinforcing user preference. Next, we formulated online learning to rank problem in bandit, with a set of users that vary in arm preference and ranking position preference. In particular, we focus on position-based user click model, which is popular in e-commerce platforms. For arm filtering, we mainly considered two strategies. One is to show users with all arms but pay incentive on agent's desired arm, thus stimulates users to pull the desired arm to some extent. The other strategy is to pick a subset of arms and show users with an ordered list of arms, on which user behavior can be seen as following position-based click model. On the other aspect, in (2), we propose policies in the face of the combinations of aforementioned user behavior. We first proposed At-Least-n Explore-Then-Commit (ALnETC) and UCB-List for incentivized bandit with self-reinforcing user preference. Then, for the previous model with delayed feedback, we improved our policy and proposed UCB-Filtering-with-Delayed-Feedback. Finally, for the problem of online learning to rank with multiple user types, we proposed two policies: GreedyRank and UCBRank, both designed for personalized treatment and equal treatment. Theoretically, we showed that all proposed policies achieve sub-linear regret, implying that all policies success in finding the optimal action in expectation. Experiments on synthetic and real-world datasets verify our theoretical results.

6.2 Limitations and Future Work

6.2.1 Fairness and Social Welfare Issues

Fairness and social welfare are always hot social issues. On the one hand, fairness is required legally for many service providers to make recommendations, in the consideration of races, genders, religions, etc; on the other hand, many recommendation systems set their goals to maximize social welfare or regard it as a constraint when maximizing profit. In Chapter 3 and 4, we proposed policies that converge the recommendation to the optimal arm over time. In the propose model, our policies achieve optimal performance, while always recommending one arm ignores the diversity of user types in a system, which can lead to monopoly of one user type in the system, thus the system is lack of diversity and fairness. In the consideration of fairness and social welfare, an improved model allows a variety of user types, each of which could behave different in arm preference. However, in such model, the perviously shown property of monopoly no longer stands, therefore, this could be a future direction for policy design.

6.2.2 General User Click Models

Click models are crucial in problem formulation, and policy design. A good click model describes user behavior accurately, and helps in designing policies that perform efficiently in practice. However, it is well-known that user behavior is complicated and hard to be formulated by a detailed model. In recommendation systems, users usually have behavior that falls in the intersection of several click models. This requires a problem formulation that models user behavior in a general click model.

In Chapter 5, we focused on position-based model, which can be naive in some real-world situations. An improved model incorporates a general click model, that allows some uncertainty on user click behavior. However, a general click model also brings challenges. For example, the uncertainty of click model makes the estimation of click model parameters much harder, leading to a much harder policy design. In future work, general click model can be a challenge, but could also shed lights in user behavior modeling.

References

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

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26.

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

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2017). Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pages 76–78. PMLR.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019). Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485.

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

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

Avadhanula, V. (2019). *The MNL-Bandit Problem: Theory and Applications*. Columbia University.

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

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

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

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

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

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

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

Chakrabarti, S., Frieze, A., and Vera, J. (2006). The influence of search engines on preferential attachment. *Internet Mathematics*, 3(3):361–381.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016a). Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016b). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778.

Chen, X. and Wang, Y. (2017). A note on a tight lower bound for mnl-bandit assortment selection models. *arXiv preprint arXiv:1709.06109*.

Chuklin, A., Markov, I., and De Rijke, M. (2022). *Click models for web search*. Springer Nature.

Chuklin, A., Markov, I., and Rijke, M. d. (2015). Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3):1–115.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94.

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

Dong, K., Li, Y., Zhang, Q., and Zhou, Y. (2020). Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615. PMLR.

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

Eick, S. G. (1988). The two-armed bandit with delayed responses. *The Annals of Statistics*, pages 254–264.

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

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

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

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

Hotchkiss, G., Alston, S., and Edwards, G. (2005). Eye tracking study.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR.

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

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

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28.

Lagrée, P., Vernade, C., and Cappe, O. (2016). Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29.

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

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. (2021). Stochastic multi-armed bandits with unrestricted delay distributions. *arXiv preprint arXiv:2106.02436*.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. (2018). Toprank: A practical algorithm for online stochastic ranking. *Advances in Neural Information Processing Systems*, 31.

Manegueu, A. G., Vernade, C., Carpentier, A., and Valko, M. (2020). Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR.

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

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

Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

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

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

Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791.

Ramezani, S. and Endriss, U. (2009). Nash social welfare in multiagent resource allocation. In *Agent-mediated electronic commerce. Designing trading strategies and mechanisms for electronic markets*, pages 117–131. Springer.

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

Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530.

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

Sierksma, G. and Zwols, Y. (2015). *Linear and integer optimization: theory and practice*. CRC Press.

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

Sorokina, D. and Cantu-Paz, E. (2016). Amazon search: The joy of ranking products. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 459–460.

Valizadegan, H., Jin, R., Zhang, R., and Mao, J. (2009). Learning to rank by optimizing ndcg measure. *Advances in neural information processing systems*, 22.

Vernade, C., Cappé, O., and Perchet, V. (2017). Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*.

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR.

Wang, S. and Huang, L. (2018). Multi-armed bandits with compensation. *Advances in Neural Information Processing Systems*, 31.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Zhou, T., Liu, J., Dong, C., and Deng, J. (2021). Incentivized bandit learning with self-reinforcing user preferences. *arXiv preprint arXiv:2105.08869*.

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

Zoghi, M., Tunys, T., Ghavamzadeh, M., Kveton, B., Szepesvari, C., and Wen, Z. (2017). Online learning to rank in stochastic click models. In *International Conference on Machine Learning*, pages 4199–4208. PMLR.

Appendix A

Proofs of Results in Chapter 3

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

Proof. 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 A.1. Next we will construct a sequence that has the same conditional distribution as $\{\chi_j\}$.

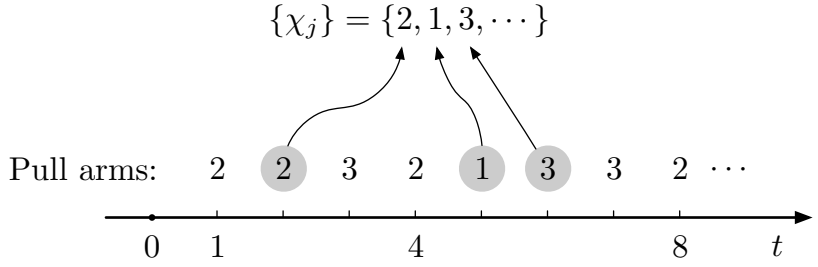


Figure A.1: This figure shows an instance of sequence $\{\chi_j\}$. At time step $t = 1$, arm 2 is pulled and generates 0 reward. At time step $t = 2$, arm 2 is pulled and generates a unit reward. Thus, the first element χ_1 in $\{\chi_j\}$ is the arm index 2 that generates the first unit reward. The 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 $B_i \coloneqq \{\sum_{k=0}^{n} r_i(k)\}_{n=0}^{\infty}$, 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 16. *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 and Khanin (2001) and Oliveira (2009) for further details. □

Proof of Lemma 16

Proof. 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,

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

□

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 $O(\log T)$.*

Proof. 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:

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

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 (A.1). Since time τ_s is defined as the earliest time to reach dominance, we can upper bound $\mathbb{E}[\tau_s - \tau_n]$ by

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

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

Lemma 17. *In ALnETC, the expected exploration phase duration $\mathbb{E}[\tau_n]$ is upper bounded by $O(\log T)$.*

□

Proof of Lemma 17

Proof. 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

$$\mathbb{E}[\tau_n] \leq n \cdot \sum_{i \in A} \frac{1+G(b,t)}{G(b,t)\mu_i} = O(n) = O(\log T). \tag{A.3}$$

Lastly, it follows from Lemma 17 that $\mathbb{E}[\tau_s] = \mathbb{E}[\tau_n] + \mathbb{E}[\tau_s - \tau_n] = O(\log T)$. This completes the proof. □

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:*

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

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:*

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

Proof. 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 18. *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 18. 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:

$$\begin{aligned}
\mathbb{E}[R_T \mid \hat{a}^* = a^*] &= \mathbb{E}[\Gamma_T^*] - \mathbb{E}[\Gamma_T \mid \hat{a}^* = a^*] \\
&= \mu^* \cdot T - \mathbb{E}[\Gamma_T \mid \hat{a}^* = a^*] \\
&= \mu^* \cdot T - \big(\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*] + \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*]\big). \qquad \text{(A.4)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] &= \mathbb{E}\Big[\sum_{i\in A}\big(n + (S_i(\tau_n) - n)\big)\Big] \\
&= m \cdot n + \mathbb{E}\Big[\sum_{i\in A}\big(T_i(\tau_n)\cdot\mu_i - n\big)\Big] \\
&= m \cdot n + \sum_{i\in A}\mu_i\left(\mathbb{E}[T_i(\tau_n)] - \frac{n}{\mu_i}\right) \\
&\geq m \cdot n + \mu_{min}\cdot\sum_{i\in A}\left(\mathbb{E}[T_i(\tau_n)] - \frac{n}{\mu_i}\right) \\
&= m \cdot n + \left(\tau_n\cdot\mu_{min} - \mu_{min}\cdot\sum_{i\in A}\frac{n}{\mu_i}\right) \\
&= \tau_n\cdot\mu_{min} + n\cdot\sum_{i\in A}\frac{\mu_i - \mu_{min}}{\mu_i}. \qquad \text{(A.5)}
\end{aligned}$$

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 (A.5) we obtain:

$$
\begin{aligned}
&\mathbb{E}[\Gamma_{\tau_s} \mid \hat{a}^* = a^*, \tau_n, \tau_s] \\
&= \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{\lambda_{a^*}(t) + G(b,t)}{1+G(b,t)} \cdot \mu^* + \sum_{i \in A} \frac{\lambda_i(t)}{1+G(b,t)} \cdot \mu_i\right] \\
&\geq \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{\lambda_{a^*}(t) + G(b,t)}{1+G(b,t)} \cdot \mu^* + \frac{(1-\lambda_{a^*}(t))}{1+G(b,t)} \cdot \mu_{min}\right] \\
&= \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \sum_{t=\tau_n+1}^{\tau_s} \mathbb{E}\left[\frac{G(b,t)}{1+G(b,t)} \cdot \mu^* + \frac{\mu_{min}}{1+G(b,t)} + \frac{\lambda_{a^*}(t)\Delta_{max}}{1+G(b,t)}\right] \\
&\geq \mathbb{E}[\Gamma_{\tau_n} \mid \tau_n] + \frac{\mu^*(\tau_s - \tau_n)G(b,t)}{1+G(b,t)} + \frac{(\tau_s - \tau_n)\mu_{min}}{1+G(b,t)} + \frac{(\tau_s - \tau_n)L_{a^*}\Delta_{max}}{1+G(b,t)} \\
&\overset{\text{(i)}}{\geq} \tau_n \mu_{min} + n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^*(\tau_s - \tau_n)G(b,t)}{1+G(b,t)} + \frac{(\tau_s - \tau_n)\mu_{min}}{1+G(b,t)} + \frac{(\tau_s - \tau_n)L_{a^*}\Delta_{max}}{1+G(b,t)} \\
&= n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1+G(b,t)} \tau_s + \tau_n \cdot \mu_{min} - \\
&\quad \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1+G(b,t)} \tau_n \\
&= n \sum_{i \in A} \frac{\mu_i - \mu_{min}}{\mu_i} + \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1+G(b,t)} \tau_s - \frac{\big(G(b,t) + L_{a^*}\big)\Delta_{max}}{1+G(b,t)} \tau_n,
\end{aligned}
\tag{A.6}
$$

where (i) is obtained by replacing $\mathbb{E}[\Gamma_{\tau_n} \mid \tau_n]$ using (A.5). Then replacing (A.4) using (A.6) and taking expectation with respect to τ_n and τ_s, we obtain:

$$
\begin{aligned}
&\mathbb{E}[R_T \mid \hat{a}^* = a^*] \\
&\leq \mu^* T - \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1+G(b,t)}\mathbb{E}[\tau_s] + \frac{\big(G(b,t)+L_{a^*}\big)\Delta_{max}}{1+G(b,t)}\mathbb{E}[\tau_n] - \\
&\quad n\sum_{i\in A}\frac{\mu_i - \mu_{min}}{\mu_i} - \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*] \\
&= \mu^*\mathbb{E}[\tau_s] - \frac{\mu^* G(b,t) + \mu_{min} + L_{a^*}\Delta_{max}}{1+G(b,t)}\mathbb{E}[\tau_s] + \frac{\big(G(b,t)+L_{a^*}\big)\Delta_{max}}{1+G(b,t)}\mathbb{E}[\tau_n] - \\
&\quad n\sum_{i\in A}\frac{\mu_i - \mu_{min}}{\mu_i} + \mu^*\big(T - \mathbb{E}[\tau_s]\big) - \mathbb{E}[\Gamma_T - \Gamma_{\tau_s} \mid \hat{a}^* = a^*] \\
&= \frac{\Delta_{max}(1-L_{a^*})}{1+G(b,t)}\mathbb{E}[\tau_s - \tau_n] + \Delta_{max}\mathbb{E}[\tau_n] - n\sum_{i\in A}\frac{\mu_i - \mu_{min}}{\mu_i} + \mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*].
\end{aligned}
\tag{A.7}
$$

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 (A.7) that

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

where (i) follows because $G(b,t) + 1 < 2G(b,t)$. By leveraging Eqs (A.3) and (A.2), 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:

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

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 19. *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:*

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

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 19, 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 $e^{-(u_0 n_0 + 1)^\gamma}$. Thus, we can upper bound the expected regret $\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*]$ by

$$\begin{aligned}
\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*] &\leq e^{-(u_0 n_0)^\gamma} + e^{-(u_0 n_0 + 1)^\gamma} + \cdots \\
&\leq \int_{u_0 n_0 - 1}^{\infty} e^{-n^\gamma} dn \\
&= C e^{-(u_0 n_0 - 1)^\gamma}, \qquad \text{(A.8)}
\end{aligned}$$

where C only depends on $u_0 n_0$ and γ such that $C = O\big((u_0 n_0)^{1-\gamma}\big)$ with $\gamma \in (0, 1/4)$. Thus Eq. (A.8) 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. □

Proof of Lemma 18

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 $-\infty < a \leq b < \infty$. Then for all $s \geq 0$*

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

Proof. Let sequences $\{X_i(t)\}$ denote the Bernoulli reward with support $\{0,1\}$ generated by arm $i \neq a^*$ 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 $S_i(\tau_n) \geq n$. By Chernoff-Hoeffding bound, at time step τ_n for arm i, we get the following:

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

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

Proof of Lemma 19

Proof. 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 19, we use the following lemma

Lemma 20. *There exists a constant n_0 such that for all $n > n_0$,*

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

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

$$\left|\frac{\sum_{j=u_0n_0}^{\infty} r_2(j)}{\mathbb{E}\big[\sum_{j=u_0n_0}^{\infty} r_2(j)\big]} - 1\right| \le (u_0n_0)^{-\frac{1}{4}}, \tag{A.9}$$

$$\forall n > n_0, \left|\frac{\sum_{j=un}^{\infty} r_2(j)}{\mathbb{E}\big[\sum_{j=un}^{\infty} r_2(j)\big]} - 1\right| \le (un)^{-\frac{1}{4}}, \tag{A.10}$$

$$\left|\frac{\sum_{j=(1-u_0)n_0}^{\infty} r_1(j)}{\mathbb{E}\big[\sum_{j=(1-u_0)n_0}^{\infty} r_1(j)\big]} - 1\right| \le \big((1-u_0)n_0\big)^{-\frac{1}{4}}, \tag{A.11}$$

$$\forall n > n_0, \left|\frac{\sum_{j=(1-u)n}^{\infty} r_1(j)}{\mathbb{E}\big[\sum_{j=(1-u)n}^{\infty} r_1(j)\big]} - 1\right| \le \big((1-u)n\big)^{-\frac{1}{4}}. \tag{A.12}$$

By Lemma 20, 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_0n_0, $E_{n_0} \cap D(u_0, n_0) = 0$, then the proof is finished since it implies

$$\mathbb{P}\bigg(\exists n > n_0, D(u_0, n_0)\bigg) \le \mathbb{P}(E_{n_0}^c) \le e^{-(u_0n_0)^{\gamma}}.$$

We consider the definition of event $D(u_0, n_0)$. By (A.9)–(A.12), we obtain

$$\begin{aligned}\sum_{i=u_0n_0}^{un-1} r_2(i) &= \sum_{i=u_0n_0}^{\infty} r_2(i) - \sum_{i=un}^{\infty} r_2(i) \\ &\ge \big(1+o(1)\big) \sum_{i=u_0n_0}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} - \big(1+o(1)\big) \sum_{i=un}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)},\end{aligned}$$

and similarly,

$$\sum_{i=n_0-u_0n_0}^{n-un-1} r_1(i) \leq \big(1+o(1)\big) \sum_{i=(1-u_0)n_0}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)} - \big(1+o(1)\big) \sum_{i=(1-u)n}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)}.$$

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

$$\begin{aligned}
&\big(1+o(1)\big) \sum_{i=u_0n_0}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} - \big(1+o(1)\big) \sum_{i=un}^{\infty} \frac{1}{\mu_2 F(i+\theta_2)} \\
&< \big(1+o(1)\big) \sum_{i=(1-u_0)n_0}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)} - \big(1+o(1)\big) \sum_{i=(1-u)n}^{\infty} \frac{1}{\mu_1 F(i+\theta_1)},
\end{aligned}$$

which implies

$$\sum_{i=u_0n_0}^{(1-u_0)n_0} \frac{1}{\mu_1 F(i+\theta_1)} < \big(1+o(1)\big) \sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(i+\theta_1)}. \tag{A.13}$$

We want to show that (A.13) cannot hold as u_0n_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

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

Also, note that $[un_0, (1-u)n_0] \subset [u_0n_0, (1-u_0)n_0]$. Therefore, there exists a constant $d \in (0,1)$ such that

$$\sum_{i=un}^{(1-u)n} \frac{1}{\mu_1 F(i+\theta_1)} \leq dk\left(\frac{n_0}{n}\right)^\alpha \sum_{i=u_0n_0}^{(1-u_0)n_0} \frac{1}{\mu_1 F(i+\theta_1)},$$

which contradicts with (A.13) since $o(1)$ goes to 0 as $u_0 n_0$ goes to infinity, and this completes the proof. □

Proof of Lemma 20

Proof. 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

$$\mathbb{P}(R_n - \mathbb{E}[R_n] > t\sqrt{Z_n}) \leq e^{-t}, \tag{A.14}$$

and

$$\mathbb{P}(R_n - \mathbb{E}[R_n] < -t\sqrt{Z_n}) \leq e^{-t}. \tag{A.15}$$

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

$$\begin{aligned}
\mathbb{P}(R_n - \mathbb{E}[R_n] > t\sqrt{Z_n}) &\overset{\text{(i)}}{=} \mathbb{P}\left(e^{s(R_n - \mathbb{E}[R_n])} > e^{st\sqrt{Z_n}}\right) \\
&\overset{\text{(ii)}}{\leq} e^{-st\sqrt{Z_n}}\mathbb{E}\left[e^{s\sum_{j\geq n}(r_i(j) - \frac{1}{h(j)})}\right] \\
&= e^{-st\sqrt{Z_n}}\prod_{j\geq n}\mathbb{E}\left[e^{s(r_i(j) - \frac{1}{h(j)})}\right] \\
&\overset{\text{(iii)}}{=} e^{-st\sqrt{Z_n}}\prod_{j\geq n}\frac{e^{-\frac{s}{h(j)}}}{1-\frac{s}{h(j)}} \\
&= e^{-st\sqrt{Z_n}}\prod_{j\geq n}e^{\frac{-s}{h(j)}}\left[1 + \frac{s}{h(j)} + \frac{\frac{s^2}{h(j)^2}}{1-\frac{s}{h(j)}}\right] \\
&\overset{\text{(iv)}}{\leq} e^{-st\sqrt{Z_n}}\prod_{j\geq n}e^{\frac{2s^2}{h(j)^2}} \\
&\leq \exp(2s^2 Z_n - st\sqrt{Z_n}), \qquad \text{(A.16)}
\end{aligned}$$

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 (A.16), 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 21 (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 21, 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 (A.14) and (A.15), and we get the inequality in Lemma 20. □

Proof of Theorem 5

Lemma 5. (UCB-Filtering-with-Delayed-Feedback) *Given a fixed time horizon T, if $G(b,t) > 1$, and $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$[1], then the pseudo regret of Algorithm 3 $\mathbb{E}[R_T]$ is upper bounded by*

$$\mathbb{E}[R_T] \leq \sum_{a \neq a^*} \frac{8\Delta_a \big(G(b,1) - 1\big) + 8\Delta^*}{\big(G(b,1) - 1\big)\Delta_a^2} \ln T + \frac{g(F,1)\Delta^* \big(\mathbb{E}[D^*(T)] + 4K\big)}{g(b,1) - 1},$$

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

$$\mathbb{E}[B_T] \leq b \cdot \frac{2G(b,1) + 1}{G(b,1) - 1} \left[\frac{8 \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \frac{8 \ln T}{\Delta_a^2} + \mathbb{E}[D^*(T)] + 4K \right].$$

[1]The notation $\Theta()$ in this work is defined as that, if $F(x) = \Theta(g(x))$, then there exist x_0 and two constants $C_1, C_2 > 0$, such that $C_1 G(x) \leq F(x) \leq C_2 g(x)$ for all $x \geq x_0$.

Proof. 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:

$$\begin{aligned}\mathbb{E}[R_T \mid \hat{a}^* = a^*] &= \mathbb{E}[R_{\tau_2} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*] \\ &= \mathbb{E}[R_{\tau_1}] + \mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{\tau_2} \mid \hat{a}^* = a^*]. \quad \text{(A.17)}\end{aligned}$$

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:

$$\begin{aligned}
&\mathbb{P}\big(\exists a \neq a^* : a \in U(t), a^* \in U(t)\big) \\
&\overset{\text{(i)}}{\leq} \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big) \cdot \mathbb{P}\big(\hat{\mu}^*(t) + c_{a^*}(t) > \hat{\mu}_a(t) - c_a(t)\big) \\
&\leq \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big) \\
&\overset{\text{(ii)}}{\leq} \mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) + \mathbb{P}\left(\hat{\mu}^*(t) - c_{a^*}(t) < \mu^* - \frac{\Delta_a}{2}\right), \qquad \text{(A.18)}
\end{aligned}$$

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. (B.4). By Chernoff-Hoeffding bound we have

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) &= \mathbb{P}\left(\hat{\mu}_a(t) - \mu_a > \frac{\Delta_a}{2} - c_a(t)\right) \\
&\leq e^{-2T_a(t)\left(\frac{\Delta_a}{2} - c_a(t)\right)^2} \\
&= e^{-\left(\ln T + \frac{\Delta_a^2}{2} T_a(t) - \Delta_a \sqrt{2T_a(t) \ln T}\right)}. \qquad \text{(A.19)}
\end{aligned}$$

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. (B.5) equals $1/T$. Note that as $T_a(t)$ increases, Eq. (B.5) decreases monotonically. Similar bound can be obtained of the second probability in Eq. (B.4). Thus, in this case, the expected

regret contributed by a suboptimal arm $a \in A$ is bounded by

$$\begin{aligned}\Delta_a \mathbb{E}[T_a(t)] &\leq \frac{8\ln T}{\Delta_a} + \Delta_a T \cdot \mathbb{P}\big(t < \tau_1 : a \in U(t), a^* \in U(t)\big) \\ &\leq \frac{8\ln T}{\Delta_a} + 2\Delta_a. \end{aligned} \tag{A.20}$$

(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

$$\Delta_a \mathbb{E}[T_a(t)] \leq \Delta_a T \cdot \mathbb{P}\big(a \in U(t), a^* \notin U(t)\big) = 2\Delta_a. \tag{A.21}$$

Summing Eq. (B.7) and Eq. (B.9) 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:

$$\mathbb{E}[\tau_1] = \sum_{a \in A} \mathbb{E}[T_a(\tau_1)] \overset{\text{(i)}}{\leq} \frac{G(b,t)+1}{G(b,t)} \left(\frac{8 \ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \left(\frac{8 \ln T}{\Delta_a^2} + 4 \right) \right), \tag{A.22}$$

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:

$$
\begin{aligned}
&\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t)] \geq \sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t)] \\
&\Rightarrow \mu_{\hat{a}^*}T_{\hat{a}^*}(\tau_1) + \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - T_{\hat{a}^*}(\tau_1)] \geq \sum_{a\neq\hat{a}^*}\mu_a T_a(\tau_1) + \sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t) - T_a(\tau_1)] \\
&\Rightarrow \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - T_{\hat{a}^*}(\tau_1)] \overset{\text{(i)}}{\geq} \sum_{a\neq\hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big) + \sum_{a\neq\hat{a}^*}\mu_a\mathbb{E}[T_a(t) - T_a(\tau_1)] \\
&\Rightarrow \frac{\mu_{\hat{a}^*}G(b,t)\mathbb{E}[t-\tau_1]}{G(b,t)+1} \overset{\text{(ii)}}{\geq} \sum_{a\neq\hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big) + \frac{\max_{a\neq\hat{a}^*}\mu_a\mathbb{E}[t-\tau_1]}{G(b,t)+1} \\
&\Rightarrow \mathbb{E}[t-\tau_1] \overset{\text{(iii)}}{\geq} \frac{G(b,t)+1}{\mu_{\hat{a}^*}G(b,t) - \max\limits_{a\neq\hat{a}^*}\mu_a}\sum_{a\neq\hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big),
\end{aligned}
\tag{A.23}
$$

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

$$
\mathbb{E}[\tau_2 - \tau_1] \leq \frac{G(b,t)+1}{\mu_{\hat{a}^*}G(b,t) - \max\limits_{a\neq\hat{a}^*}\mu_a}\sum_{a\neq\hat{a}^*}\big(\frac{8\mu_a}{\Delta_a^2}\ln T + 4\mu_a\big). \tag{A.24}
$$

Thus, we can bound the regret during the exploitation phase $\mathbb{E}[R_{\tau_2} - R_{\tau_1} \mid \hat{a}^* = a^*]$ in (A.17) 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 (B.11) and (B.13), the expected accumulative payment $\mathbb{E}[B_T]$ can also be upper bounded by

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

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. □

Appendix B

Proofs of Results in Chapter 4

Proof of Lemma 5

Lemma 5. (UCB-Filtering-with-Delayed-Feedback) *Given a fixed time horizon T, if $G(b,t) > 1$, and $F(x) = \Theta(x^\alpha)$ with $\alpha > 1$*[1]*, then the pseudo regret of Algorithm 3 $\mathbb{E}[R_T]$ is upper bounded by*

$$\mathbb{E}[R_T] \leq \sum_{a \neq a^*} \frac{8\Delta_a\big(G(b,1)-1\big) + 8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2} \ln T + \frac{g(F,1)\Delta^*\big(\mathbb{E}[D^*(T)] + 4K\big)}{g(b,1)-1},$$

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

$$\mathbb{E}[B_T] \leq b \cdot \frac{2G(b,1)+1}{G(b,1)-1}\left[\frac{8\ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \frac{8\ln T}{\Delta_a^2} + \mathbb{E}[D^*(T)] + 4K\right].$$

[1]The notation $\Theta()$ in this work is defined as that, if $F(x) = \Theta(g(x))$, then there exist x_0 and two constants $C_1, C_2 > 0$, such that $C_1 G(x) \leq F(x) \leq C_2 g(x)$ for all $x \geq x_0$.

Proof. By the law of total expectation, the expected regret up to time T can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}[R_T] \\
&= \mathbb{E}[R_{t_0}] + \mathbb{E}[R_T \mid \hat{a}^* = a^*]\mathbb{P}(\hat{a}^* = a^*) + \mathbb{E}[R_T \mid \hat{a}^* \neq a^*]\mathbb{P}(\hat{a}^* \neq a^*) \\
&\leq \mathbb{E}[R_{t_0}] + \mathbb{E}[R_T \mid \hat{a}^* = a^*] + T \cdot \mathbb{P}(\hat{a}^* \neq a^*) \\
&= \mathbb{E}[R_{t_0}] + \mathbb{E}[R_{t_2} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{t_2} \mid \hat{a}^* = a^*] + T \cdot \mathbb{P}(\hat{a}^* \neq a^*) \\
&= \mathbb{E}[R_{t_0}] + \mathbb{E}[R_{t_1}] + \mathbb{E}[R_{t_2} - R_{t_1} \mid \hat{a}^* = a^*] + \mathbb{E}[R_T - R_{t_2} \mid \hat{a}^* = a^*] + T \cdot \mathbb{P}(\hat{a}^* \neq a^*).
\end{aligned} \tag{B.1}$$

We start from bounding the term $\mathbb{E}[R_{t_0}]$. Note that the goal of the initialization step is to ensure that the confidence interval for each arm is initialized. Then we have

$$\begin{aligned}
\mathbb{E}[R_{t_0}] &\leq \mathbb{E}[t_0] \\
&\overset{\text{(i)}}{\leq} m \cdot \left(\frac{1 + G(b,1)}{G(b,1)} + \mathbb{E}[\tau_1] \right) \\
&\overset{\text{(i)}}{<} m \cdot (\mathbb{E}[\tau_1] + 2),
\end{aligned} \tag{B.2}$$

where (i) is because that by definition of initialization, for each arm a, it contributes to t_0 in terms of its incentivizing attempts for being pulled by users once, and its delay period after which arm a will be regarded as initialized, then at time step t, when offered incentive b, arm a has probability no less than $G(b,t)/\big(1+G(b,t)\big)$ to be pulled by users, thus the number of attempts for arm a to be pulled once is a geometric random variable with parameter no less than $G(b,t)/\big(1+G(b,t)\big)$, with expectation value $\big(1+G(b,t)\big)/G(b,t)$, and (ii) is because that by condition $G(b,t) > 1$, we have

$\big(1 + G(b,t)\big)/G(b,t) < 2$.

Then we bound the term $\mathbb{E}[R_{t_1}]$. During the exploration phase, since the regret is caused by pullings of sup-optimal arms, the expected regret after $t \leq t_1$ 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 $\mathbb{E}[R_{t_1}]$ by bounding each $\mathbb{E}[T_a(t_1)]$ for $a \neq a^*$. Let $\mathcal{U}(t)$ denote the set of arms that are activated, i.e., can get incentive, at time t. Consider the following two cases during the exploration phase:

(a) At time $t \leq t_1$, $a^* \in \mathcal{U}(t)$ and there exists at least one suboptimal arm $a \in \mathcal{A}, a \neq a^*$ such that $a \in \mathcal{U}(t)$. Recall that $c_a(t) = \sqrt{\ln T / 2\big(T_a(t) - D_a(t)\big)}$ is the confidence interval of arm a at time step t. In this case, we have:

$$\begin{aligned}
&\mathbb{P}\big(\exists a \neq a^* : a \in \mathcal{U}(t), a^* \in \mathcal{U}(t)\big) \qquad \text{(B.3)}\\
&\overset{\text{(i)}}{\leq} \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big) \cdot \mathbb{P}\big(\hat{\mu}^*(t) + c_{a^*}(t) > \hat{\mu}_a(t) - c_a(t)\big)\\
&\leq \mathbb{P}\big(\hat{\mu}_a(t) + c_a(t) > \hat{\mu}^*(t) - c_{a^*}(t)\big)\\
&\overset{\text{(ii)}}{\leq} \mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) + \mathbb{P}\left(\hat{\mu}^*(t) - c_{a^*}(t) < \mu^* - \frac{\Delta_a}{2}\right), \qquad \text{(B.4)}
\end{aligned}$$

where (i) is obtained since arms $a, a^* \in \mathcal{U}(t)$ implies that the upper confidence bound of either arm 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 the event $\{\hat{\mu}_a(t) + c_a(t) > \mu_a + \Delta_a/2\}$ or the event $\{\hat{\mu}^*(t) - c_{a^*}(t) < \mu^* - \Delta_a/2\}$. We consider the first probability in Eq. (B.4), and similar bound will be obtained for the

second probability. By Chernoff-Hoeffding bound we have

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_a(t) + c_a(t) > \mu_a + \frac{\Delta_a}{2}\right) &= \mathbb{P}\left(\hat{\mu}_a(t) - \mu_a > \frac{\Delta_a}{2} - c_a(t)\right) \\ &\leq e^{-2(T_a(t)-D_a(t))\left(\frac{\Delta_a}{2}-c_a(t)\right)^2} \\ &= e^{-\left(\ln T + \frac{\Delta_a^2}{2}(T_a(t)-D_a(t)) - \Delta_a\sqrt{2(T_a(t)-D_a(t))\ln T}\right)}. \quad \text{(B.5)}\end{aligned}$$

Let $\frac{\Delta_a^2}{2}(T_a(t)-D_a(t)) - \Delta_a\sqrt{2(T_a(t)-D_a(t))\ln T} = 0$, we obtain $T_a(t) = 8\ln T/\Delta_a^2 + D_a(t)$ and Eq. (B.5) equals $1/T$. Note that as $(T_a(t) - D_a(t))$ increases, Eq. (B.5) decreases monotonically. Thus, in this case, the expected regret contributed by a suboptimal arm $a \in A$ during the exploration phase is bounded by

$$\begin{aligned}\Delta_a\mathbb{E}[T_a(t_1)] &\leq \frac{8\ln T}{\Delta_a} + \Delta_a\left(\mathbb{E}[D_a(t_1)] + T\cdot\mathbb{P}\big(t < t_1 : a \in \mathcal{U}(t), a^* \in \mathcal{U}(t)\big)\right) \\ &\leq \frac{8\ln T}{\Delta_a} + \Delta_a\big(\mathbb{E}[D_a(t_1)] + 2\big) \quad \text{(B.6)} \\ &\leq \frac{8\ln T}{\Delta_a} + \Delta_a\big(\mathbb{E}[D_a^*(t_1)] + 2\big). \quad \text{(B.7)}\end{aligned}$$

(b) At time $t \leq t_1$, a^* is eliminated by some suboptimal arm $a \in \mathcal{U}(t), a \neq a^*$. In this case, with similar technique as that in case **(a)** and Chernoff-Hoeffding bound,

we have

$$
\begin{aligned}
&\mathbb{P}\big(\exists a \neq a^* : a \in \mathcal{U}(t), a^* \notin \mathcal{U}(t)\big) \\
&\leq \mathbb{P}\big(\hat{\mu}_a(t) - c_a(t) > \hat{\mu}^*(t) + c_{a^*}(t)\big) \\
&\leq \mathbb{P}\big(\hat{\mu}_{a^*}(t) + c_{a^*}(t) \leq \mu_{a^*} - \frac{\Delta_a}{2}\big) + \mathbb{P}\big(\hat{\mu}_a(t) - c_a(t) \geq \mu_a + \frac{\Delta_a}{2}\big) \\
&\leq e^{-2(T_{a^*}(t)-D_{a^*}(t))\left(\frac{\Delta_a}{2}+c_{a^*}(t)\right)^2} + e^{-2(T_a(t)-D_a(t))\left(\frac{\Delta_a}{2}+c_a(t)\right)^2} \\
&= e^{-\frac{\Delta_a^2}{2}(T_{a^*}(t)-D_{a^*}(t))-\ln T-\Delta_a\sqrt{2(T_{a^*}(t)-D_{a^*}(t))\ln T}} \\
&\quad + e^{-\frac{\Delta_a^2}{2}(T_a(t)-D_a(t))-\ln T-\Delta_a\sqrt{2(T_a(t)-D_a(t))\ln T}} \\
&\leq 2T^{-1}.
\end{aligned} \tag{B.8}
$$

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

$$
\Delta_a \mathbb{E}[T_a(t_1)] \leq \Delta_a T \cdot \mathbb{P}\big(a \in \mathcal{U}(t), a^* \notin \mathcal{U}(t)\big) = 2\Delta_a. \tag{B.9}
$$

Summing Eq. (B.7) and Eq. (B.9) over all suboptimal arms, the expected regret during the exploration phase is bounded by:

$$
\mathbb{E}[R_{t_1}] \leq \sum_{a \neq a^*} \frac{8 \ln T}{\Delta_a} + \Delta_a\big(\mathbb{E}[D_a^*(t_1)] + 4\big). \tag{B.10}
$$

During the exploration phase at time step $t < t_1$, since the agent offers incentive payment b to the user for pulling arm i, the probability that the arm i is pulled is $\big(p_i(t)+G(b,t)\big)/\big(1+G(b,t)\big)$ lower bounded by $G(b,t)/\big(1+G(b,t)\big)$. Thus, the number of attempts for arm i to get pulled is a geometric random variable with expectation

no larger than $\big(1 + G(b,t)\big)/G(b,t)$. Since the above cases **(a)** and **(b)** imply that at most $8\ln T/\Delta_a^2 + \mathbb{E}[D_a^*(t_1)] + 4$ expected number of pullings from each suboptimal arm ensures a good estimation of the optimal arm, thus, the expected number of incentivizing attempts on a suboptimal arm a to guarantee $8\ln T/\Delta_a^2 + \mathbb{E}[D_a^*(t_1)] + 4$ number of pullings is upper bounded by:

$$\mathbb{E}\Big[\sum_{s=1}^{t_1} 1_{\{I'_s=a\}}\Big] \leq \frac{G(b,1)+1}{G(b,1)}\Big(\frac{8\ln T}{\Delta_a^2} + \mathbb{E}[D_a^*(t_1)] + 4\Big).$$

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

$$\mathbb{E}[t_1] = \sum_{a\in A}\mathbb{E}\Big[\sum_{s=1}^{t_1} 1_{\{I'_s=a\}}\Big] \overset{\text{(i)}}{\leq} \frac{G(b,1)+1}{G(b,1)}\Big[\frac{8\ln T}{\Delta_{min}^2} + \sum_{a\neq a^*}\Big(\frac{8\ln T}{\Delta_a^2} + 4\Big) + \sum_{a\in A} D_a^*(t_1)\Big], \tag{B.11}$$

where (i) is due to the requirement of $T_{a^*}(t_1)$ to be at most $8\ln T/\Delta_{min}^2 + \mathbb{E}[D_{a^*}^*(t_1)]$, 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 $t \geq t_1$ if

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

We tighten the condition by narrowing the left-hand-side and amplifying the right-hand-side, and obtain the sufficient condition of dominance as follows:

$$
\begin{aligned}
&\mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - D_{\hat{a}^*}(t)] \geq \sum_{a\neq\hat{a}^*} \mu_a \mathbb{E}[T_a(t)] \\
&\Rightarrow \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - D^*_{\hat{a}^*}(t) - T_{\hat{a}^*}(t_1)] \geq \sum_{a\neq\hat{a}^*} \mu_a \bigg(\mathbb{E}[T_a(t_1)] + \mathbb{E}[T_a(t) - T_a(t_1)] \bigg) \\
&\Rightarrow \mu_{\hat{a}^*}\mathbb{E}[T_{\hat{a}^*}(t) - D^*_{\hat{a}^*}(t) - T_{\hat{a}^*}(t_1)] \overset{\text{(i)}}{\geq} \\
&\quad \sum_{a\neq\hat{a}^*} \bigg[\frac{8\mu_a}{\Delta_a^2} \ln T + \mu_a \big(\mathbb{E}[D^*_a(t)] + 4 \big) + \mu_a \mathbb{E}[T_a(t) - T_a(t_1)] \bigg] \\
&\Rightarrow \mu_{\hat{a}^*} \bigg[\frac{G(b,1)\mathbb{E}[t-\tau_1]}{G(b,1)+1} - \mathbb{E}[D^*_{\hat{a}^*}(t)] \bigg] \overset{\text{(ii)}}{\geq} \\
&\quad \sum_{a\neq\hat{a}^*} \bigg[\frac{8\mu_a}{\Delta_a^2} \ln T + \mu_a \big(\mathbb{E}[D^*_a(t)] + 4 \big) \bigg] + \frac{\mu_{\hat{a}^*}\mathbb{E}[t-\tau_1]}{G(b,1)+1} \\
&\Rightarrow \mathbb{E}[t-\tau_1] \overset{\text{(iii)}}{\geq} \frac{G(b,1)+1}{G(b,1)-1} \bigg[\sum_{a\neq\hat{a}^*} \frac{8\ln T}{\Delta_a^2} + \sum_{a\in A} \big(\mathbb{E}[D^*_a(t)] + 4 \big) \bigg],
\end{aligned}
\tag{B.12}
$$

where (i) is obtained since $T_{\hat{a}^*}(t_1) > 0$ and $D^*_{\hat{a}^*}(t) \geq D_{\hat{a}^*}(t)$, (ii) is because by incentivizing arm $\hat{a}^*$, we have $\hat{p}_{\hat{a}^*}(t) \geq G(b,t)/\big(G(b,t)+1\big)$ and $\hat{p}_a(t) \leq 1/\big(G(b,t)+1\big)$ for $a \neq \hat{a}^*$, and (iii) is the rearrangement. Since time t_2 is defined as the earliest time to reach dominance, we can upper bound $\mathbb{E}[t_2 - t_1]$ by

$$
\mathbb{E}[t_2 - t_1] \leq \frac{G(b,1)+1}{G(b,1)-1} \bigg[\sum_{a\neq\hat{a}^*} \frac{8\ln T}{\Delta_a^2} + \sum_{a\in A} \big(\mathbb{E}[D^*_a(t_2)] + 4 \big) \bigg]. \tag{B.13}
$$

Thus, we can bound the regret during the exploitation phase $\mathbb{E}[R_{t_2} - R_{t_1} \mid \hat{a}^* = a^*]$ in (B.1) by

$$\begin{aligned}\mathbb{E}[R_{t_2} - R_{t_1} \mid \hat{a}^* = a^*] &\overset{\text{(i)}}{\leq} \frac{\Delta^*}{G(b,1)+1} \cdot \mathbb{E}[t_2 - t_1] \\ &\leq \sum_{a \neq a^*} \frac{8\Delta^* \ln T}{\Delta_a^2 (G(b,1)-1)} + \sum_{a \in A} \frac{\Delta^* \big(\mathbb{E}[D_a^*(t_2)] + 4\big)}{G(b,1)-1}, \end{aligned} \tag{B.14}$$

where (i) follows because during the exploitation phase there is always a positive probability $\hat{p}_a(t)$ which is at most $1/\big(G(b,t)+1\big)$ to pull suboptimal arm a. After arm $\hat{a}^*$ dominates, we want to prove that arm $\hat{a}^*$ has exponentially large probability to achieve monopoly, so as to upper bound the regret in the self-sustaining phase $\mathbb{E}[R_T - R_{t_2} \mid \hat{a}^* = a^*]$. Since the dominance proof only involves the accumulated reward $S_a(t)$, our situation reduces to the non-delay case in expectation. Thus, we omit the proof in Zhou et al. (2021) and show the result by

$$\mathbb{E}[R_T - R_{\tau_s} \mid \hat{a}^* = a^*] \leq e^{-(\log T)^\gamma} + e^{-(\log T + 1)^\gamma} + \cdots, \tag{B.15}$$

with the summation on the right hand side bounded by $O\big((\log T)^{1-\gamma} e^{-(\log T)^\gamma}\big)$ and $\gamma \in (0, 1/4)$. Now, summing up Eqs. (B.2),(B.8),(B.10),(B.14),(B.15), we obtain the regret upper bound stated in Lemma 5.

By using Eqs (B.11) and (B.13), the expected incentive $\mathbb{E}[B_T]$ can also be upper bounded by

$$
\begin{aligned}
&\mathbb{E}[B_T] \\
&= (\mathbb{E}[t_1] + \mathbb{E}[t_s - t_1]) \cdot b \\
&\leq \frac{G(b,1)+1}{G(b,1)}\Big[\frac{8b\ln T}{\Delta_{min}^2} + \sum_{a\neq a^*}\big(\frac{8b\ln T}{\Delta_a^2} + 4b\big) + \mathbb{E}[D^*(t_1)]\Big] + \\
&\quad \frac{G(b,1)+1}{G(b,1)-1}\Big[\sum_{a\neq \hat{a}^*}\frac{8b\ln T}{\Delta_a^2} + b\mathbb{E}[D^*(t_2)] + 4bK\Big] \\
&\overset{\text{(i)}}{\leq} \frac{G(b,1)+1}{G(b,1)}\Big[\frac{8b\ln T}{\Delta_{min}^2} + \sum_{a\neq a^*}\frac{8b\ln T}{\Delta_a^2} + b\mathbb{E}[D^*(t_2)] + 4bK\Big] + \\
&\quad \frac{G(b,1)+1}{G(b,1)-1}\Big[\sum_{a\neq \hat{a}^*}\frac{8b\ln T}{\Delta_a^2} + b\mathbb{E}[D^*(t_2)] + 4bK\Big] \\
&= \frac{G(b,1)+1}{G(b,1)}\cdot\frac{8b\ln T}{\Delta_{min}^2} + \\
&\quad \Big(\frac{G(b,1)+1}{G(b,1)} + \frac{G(b,1)+1}{G(b,1)-1}\Big)\cdot\Big(\sum_{a\neq \hat{a}^*}\frac{8b\ln T}{\Delta_a^2} + b\mathbb{E}[D^*(t_2)] + 4bK\Big) \\
&\overset{\text{(ii)}}{\leq} \frac{2G(b,1)+1}{G(b,1)-1}\Big[\frac{8b\ln T}{\Delta_{min}^2} + \sum_{a\neq a^*}\frac{8b\ln T}{\Delta_a^2} + b\mathbb{E}[D^*(t_2)] + 4bK\Big],
\end{aligned}
$$

where (i) follows from $D^*(t_2) \geq D^*(t_1)$ since $t_2 \geq t_1$, and (ii) follows from rearranging of the coefficients containing $G(b,1)$. □

Proof of Theorem 7

Theorem 7. (Arm-Independent Delay) *Under i.i.d. delays with a finite expectation and the conditions of Lemma 5, the pseudo regret of Algorithm 3* $\mathbb{E}[R_T]$ *is upper bounded by*

$$\left[\frac{2G(b,1)\Delta^*}{G(b,1)-1}+\sum_{a\neq a^*}\frac{8\Delta_a\big(G(b,1)-1\big)+8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2}\right]\ln T+$$
$$\frac{G(b,1)\Delta^*\big(\sqrt{4\mathbb{E}[\tau_1]\ln T}+\mathbb{E}[\tau_1]+4K+1\big)}{G(b,1)-1},$$

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

$$b\cdot\frac{2G(b,1)+1}{G(b,1)-1}\left[\left(2+\frac{8}{\Delta_{min}^2}+\sum_{a\neq a^*}\frac{8}{\Delta_a^2}\right)\ln T+\sqrt{4\mathbb{E}[\tau_1]\ln T}+\mathbb{E}[\tau_1]+4K+1\right].$$

Proof. Combining the results in Lemma 5 and Lemma 6, we obtain that the regret is upper bounded by

$$\mathbb{E}[R_T]$$
$$\leq\sum_{a\neq a^*}\frac{8\Delta_a\big(G(b,1)-1\big)+8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2}\ln T+\frac{G(b,1)\Delta^*\big(\mathbb{E}[\tau_1]+2\ln T+\sqrt{4\mathbb{E}[\tau_1]\ln T}+4K+1\big)}{G(b,1)-1}.$$

Then by some straightforward rearrangements we obtain the regret upper bound in Theorem 7. Similarly, by replacing $\mathbb{E}[D^*(T)]$ with its upper bound in the incentive upper bound, we obtain the rearranged incentive upper bound in Theorem 7. □

Proof of Lemma 8

Lemma 8. *Under Assumption 1, given a finite number of arms $K > 0$, it holds that*

$$\mathbb{E}[D^*(t)] \leq \sum_{a \in \mathcal{A}} 2\mathbb{E}[\tau_{a,1}] + 3K \log \frac{t}{K}.$$

Proof. We consider a two-armed model with arm a and b, with arm-dependent delay distribution $\mathcal{T}_a$ and $\mathcal{T}_b$. Without loss of generality, we assume that $\mathbb{E}[\tau_{a,1}] \geq \mathbb{E}[\tau_{b,1}]$. Then, the worst case where $D^*(t)$ is maximized is that one first consecutively pull arm a for t_a times, then consecutively pull arm b for the rest $t - t_a$ times. Now, consider another two cases where arm a and arm b are pulled for t_a times and $t - t_a$ times independently, we denote their maximum outstanding feedback up to time step t_a and $t - t_a$ by $D_a^*(t_a)$ and $D_b^*(t - t_a)$ respectively. We obtain the relationship below at time step t:

$$D^*(t) \leq D_a^*(t_a) + D_b^*(t - t_a),$$

Now, by the result in Lemma 6, we obtain that

$$
\begin{aligned}
&\mathbb{E}[D^*(t)] \\
&\leq \mathbb{E}[D_a^*(t_a)] + \mathbb{E}[D_b^*(t-t_a)] \\
&\overset{\text{(i)}}{\leq} \sum_{i=a,b} \mathbb{E}[\tau_{i,1}] + 2\big(\log t_a + \log(t-t_a)\big) + 2\sqrt{\mathbb{E}[\tau_{a,1}]\log t_a} + 2\sqrt{\mathbb{E}[\tau_{b,1}]\log(t-t_a)} \\
&\overset{\text{(ii)}}{\leq} \sum_{i=a,b} \mathbb{E}[\tau_{i,1}] + 2\big(\log t_a + \log(t-t_a)\big) + \mathbb{E}[\tau_{a,1}] + \log t_a + \mathbb{E}[\tau_{b,1}] + \log(t-t_a) \\
&= \sum_{i=a,b} 2\mathbb{E}[\tau_{i,1}] + 3\big(\log t_a + \log(t-t_a)\big) \\
&\overset{\text{(iii)}}{\leq} \sum_{i=a,b} 2\mathbb{E}[\tau_{i,1}] + 3 \cdot 2\log\frac{t}{2},
\end{aligned}
$$

where (i) is by the result in Lemma 6; (ii) comes from the inequality $2\sqrt{ab} \leq a+b$; and (iii) comes from the fact that $t/2$ is the maximizer of function $h(x) = \log x + \log(t-x)$. This result can be straightforwardly extended to multi-armed model with $K \geq 2$ by the relationship $D^*(t) \leq \sum_{i\in\mathcal{A}} D_i^*(t_i)$ with $\sum_{i\in\mathcal{A}} t_i = t$. □

Proof of Theorem 9

Theorem 9. (Arm-Dependent Delay) *Under Assumption 1 and the conditions of Lemma 5, the pseudo regret of Algorithm 3* $\mathbb{E}[R_T]$ *is upper bounded by*

$$
\sum_{a\neq a^*} \frac{8\Delta_a\big(G(b,1)-1\big) + 8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2} \ln T + \frac{G(b,1)\Delta^*\big(3K\ln\frac{T}{K} + \sum_{a\in\mathcal{A}} 2\mathbb{E}[\tau_{a,1}] + 4K\big)}{G(b,1)-1},
$$

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

$$b \cdot \frac{2G(b,1)+1}{G(b,1)-1}\left[\left(\frac{8}{\Delta_{min}^2}+\sum_{a\neq a^*}\frac{8}{\Delta_a^2}\right)\ln T + 3K\ln\frac{T}{K}+\sum_{a\in\mathcal{A}}\mathbb{E}[\tau_{a,1}]+4K\right].$$

Proof. Combining the results in Lemma 5 and Lemma 8, we obtain that the regret is upper bounded by

$$\begin{aligned}&\mathbb{E}[R_T]\\ &\leq \sum_{a\neq a^*}\frac{8\Delta_a\big(G(b,1)-1\big)+8\Delta^*}{\big(G(b,1)-1\big)\Delta_a^2}\ln T + \frac{G(b,1)\Delta^*\big(\sum_{a\in A}2\mathbb{E}[\tau_{a,1}]+3K\log\frac{t}{K}+4K\big)}{G(b,1)-1}.\end{aligned}$$

Thus we obtain the regret upper bound in Theorem 9. Similarly, by replacing $\mathbb{E}[D^*(T)]$ with its upper bound in the incentive upper bound, we obtain the rearranged incentive upper bound in Theorem 9. □

Appendix C

Proofs of Results in Chapter 5

Proof of Lemma 10

Lemma 10. For each user type $i \in [N]$ and position $k \in [K]$, for any constant $\epsilon \geq 0$, the position preference estimator $E(\boldsymbol{T}(t), \boldsymbol{S}(t))$ achieves a concentration bound as follows:

$$\mathbb{P}\left(|\hat{\rho}_{i,k}(t) - \rho_{i,k}| \geq \max_{j \in [M]} \sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}} \right) \leq MKt^{-2\epsilon}. \tag{5.1}$$

Proof. In Algorithm 4, we denote a random vector $\boldsymbol{v}_{i,j}(t) = \boldsymbol{S}_{i,j}(t)/\boldsymbol{T}_{i,j}(t)$ for user i and arm j. Given that each entry of vectors $\boldsymbol{S}_{i,j}(t)$ and $\boldsymbol{T}_{i,j}(t)$ is unbiased, and the arriving user at time t views exactly one position, for any time step t with arriving user $I(t) = i$ and arm $j \in m(t)$, we have $\mathbb{E}[v_{i,j,k}(t) - v_{i,j,k}(t-1)] = \mu_{i,j}\rho_{i,k}$ for position k. Then, by Hoeffding's Inequality, for any $\epsilon \geq 0$ we have

$$\mathbb{P}\left(|v_{i,j,k}(t) - \mu_{i,j}\rho_{i,k}| \geq \sqrt{\frac{\epsilon \ln t}{T_{i,j,k}(t)}} \right) \leq t^{-2\epsilon}, i \in [N], j \in [M], k \in [K]. \tag{C.1}$$

Summing $v_{i,j,k}(t)$ in (C.1) over k, by union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k\in[K]}\left(|v_{i,j,k}(t)-\mu_{i,j}\rho_{i,k}|\right)\geq K\sqrt{\frac{\epsilon\ln t}{T_{i,j,k}(t)}}\right)\\
=\ &\mathbb{P}\left(|\boldsymbol{v}_{i,j}(t)-\mu_{i,j}|\geq K\sqrt{\frac{\epsilon\ln t}{T_{i,j,k}(t)}}\right)\\
\leq\ &Kt^{-2\epsilon}.
\end{aligned}$$

To prove Eq. (10), we start from one direction of the inequality. By setting $\epsilon=0$, it is obvious that $\mathbb{P}(\boldsymbol{v}_{i,j}(t)\leq\mu_{i,j})\leq 1$. Then, for any position $k\in[K]$, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\boldsymbol{v}_{i,j}(t)}\geq\frac{1}{\mu_{i,j}},\ v_{i,j,k}(t)\geq\mu_{i,j}\rho_{i,k}+\sqrt{\frac{\epsilon\ln t}{T_{i,j,k}(t)}}\right)\\
=\ &\mathbb{P}\left(\frac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)}\geq\rho_{i,k}+\sqrt{\frac{\epsilon\ln t}{\mu_{i,j}^2T_{i,j,k}(t)}}\right)\\
\leq\ &Kt^{-2\epsilon}.
\end{aligned}$$

Averaging the term $\frac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)}$ over arms $j \in [M]$, by union bound we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{M}\sum_{j\in[M]}\frac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)} \geq \rho_{i,k} + \frac{1}{M}\sum_{j\in[M]}\sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \leq MKt^{-2\epsilon}, \\
\Rightarrow &\mathbb{P}\left(\hat{\rho}_{i,k} \geq \rho_{i,k} + \frac{1}{M}\sum_{j\in[M]}\sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \leq MKt^{-2\epsilon}, \\
\Rightarrow &\mathbb{P}\left(\hat{\rho}_{i,k} \geq \rho_{i,k} + \max_{j\in[M]}\sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \leq MKt^{-2\epsilon}.
\end{aligned}$$

Thus, we obtain one direction of Eq. (10). Similarly, to prove the reversed direction, for any position $k \in [K]$, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\boldsymbol{v}_{i,j}(t)} \leq \frac{1}{\mu_{i,j}},\ v_{i,j,k}(t) \leq \mu_{i,j}\rho_{i,k} - \sqrt{\frac{\epsilon \ln t}{T_{i,j,k}(t)}}\right) \\
&= \mathbb{P}\left(\frac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)} \leq \rho_{i,k} - \sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \\
&\leq Kt^{-2\epsilon}.
\end{aligned}$$

Averaging the term $\dfrac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)}$ over arms $j \in [M]$, we obtain the following

$$\mathbb{P}\left(\frac{1}{M}\sum_{j\in[M]}\frac{v_{i,j,k}(t)}{\boldsymbol{v}_{i,j}(t)} \le \rho_{i,k} - \frac{1}{M}\sum_{j\in[M]}\sqrt{\frac{\epsilon\ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \le MKt^{-2\epsilon},$$

$$\Rightarrow \mathbb{P}\left(\hat{\rho}_{i,k} \le \rho_{i,k} - \frac{1}{M}\sum_{j\in[M]}\sqrt{\frac{\epsilon\ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \le MKt^{-2\epsilon},$$

$$\Rightarrow \mathbb{P}\left(\hat{\rho}_{i,k} \le \rho_{i,k} - \max_{j\in[M]}\sqrt{\frac{\epsilon\ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}\right) \le MKt^{-2\epsilon}.$$

Thus, we obtain both directions of Eq. (10). □

Proof of Lemma 11

Lemma 11. For user type i and arm j, denote the unbiased empirical arm means by $\hat{\mu}_{ij}(t) = \|\boldsymbol{S}_{i,j}(t-1)\|_1 / N_{ij}(t-1)$. Then, conditioned on event $\mathcal{N}_t$ and for any $\epsilon \ge 0$ we have

$$\mathbb{P}\left(\left|\hat{\mu}_{i,j}(t) - \mu_{ij}\right| \ge \frac{\sqrt{2\epsilon\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{i,j}(t)} \;\middle|\; \mathcal{N}_t\right) \le \epsilon e^{1-\epsilon}\log t. \tag{5.2}$$

We will leverage the following proposition in the proof.

Proposition 1 (Lagrée et al. (2016), Proposition 8)**.** *Given user $i \in [N]$ and arm $j \in [M]$, for any $\epsilon \geq 0$, we have*

$$\mathbb{P}\left(\left| \bar{\mu}_{ij}(t) - \mu_{ij} \right| \geq \frac{\sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{\bar{N}_{ij}(t)} \right) \leq \epsilon e^{1-\epsilon} \log t. \tag{C.2}$$

Proof. Assume that the position probabilities $\boldsymbol{\rho}$ are known, then we can replace Line 16 in Algorithm 5 by $\bar{N}_{i,j}(t) \leftarrow \bar{N}_{i,j}(t-1) + \rho_{i,\sigma_t(j)}$, and denote unbiased empirical means by $\bar{\mu}_{i,j}(t) = \sum_k S_{i,j,k}(t-1)/\bar{N}_{i,j}(t-1)$. Formally, define $\bar{N}_{i,j}(t) = \sum_{s=1}^{t} \rho_{i,\sigma_{s-1}(j)} \cdot 1\{j \in \mathrm{M}_{\sigma_s}\}$, and define $N_{i,j}(t) = \sum_{s=1}^{t} \hat{\rho}_{i,\sigma_{s-1}(j)} \cdot 1\{j \in \mathrm{M}_{\sigma_s}\}$, where $1\{X\}$ is an indicator of event X. Then we have:

$$|N_{i,j}(t) - \bar{N}_{i,j}(t)|/\|\boldsymbol{T}_{i,j}(t)\|_1 = \frac{1}{\|\boldsymbol{T}_{i,j}(t)\|_1} \sum_{s=1}^{t} \left[\left| \hat{\rho}_{i,\sigma_{s-1}(j)} - \rho_{i,\sigma_{s-1}(j)} \right| \cdot 1\{j \in \mathrm{M}_{\sigma_s}\} \right].$$

Define a "good" event $\mathcal{N}_t$ as follows: at time t, for any user $i \in [N]$ and any arm $j \in [M]$, there exists $\epsilon \geq 0$ such that

$$|N_{i,j}(t) - \bar{N}_{i,j}(t)| < \|\boldsymbol{T}_{i,j}(t)\|_1 \max_{j \in [M]} \sqrt{\epsilon \ln t / \left(\mu_{i,j}^2 T_{i,j,k}(t)\right)}.$$

Combining with Lemma 10, we obtain that $\mathbb{P}\left(\mathcal{N}_t^C\right)$ is upper bounded by

$$\frac{1}{\|\boldsymbol{T}_{i,j}(t)\|_1} \sum_{s=1}^{t} \left[1\{j \in \mathrm{M}_{\sigma_s}\} \cdot \mathbb{P}\left(\left| \hat{\rho}_{i,\sigma_{s-1}(j)} - \rho_{i,\sigma_{s-1}(j)} \right| \geq \max_{j \in [M]} \sqrt{\frac{\epsilon \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}} \right) \right]$$

$$\leq MKt^{-2\epsilon}.$$

Then, with Proposition 1, we have

$$
\begin{aligned}
&\mathbb{P}\left(\bar{\mu}_{ij}(t) - \mu_{ij} \geq \frac{\sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{\bar{N}_{ij}(t)}\right) \\
&= \mathbb{P}\left(\hat{\mu}_{ij}(t) - \frac{\mu_{ij}\bar{N}_{ij}(t)}{N_{i,j}(t)} \geq \frac{\sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{ij}(t)}\right) \\
&= \mathbb{P}\left(\hat{\mu}_{ij}(t) - \mu_{ij} \geq \frac{\mu_{i,j}(\bar{N}_{i,j}(t) - N_{i,j}(t)) + \sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{ij}(t)}\right) \\
&\leq \epsilon e^{1-\epsilon}\log t.
\end{aligned}
$$

Replacing $\left(\bar{N}_{i,j}(t) - N_{i,j}(t)\right)$ by $\|\boldsymbol{T}_{i,j}(t)\|_1 \max_{j\in[M]}\sqrt{\frac{\epsilon_1 \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}$ with $\epsilon_1 = \epsilon/\ln t$, conditioned on $\mathcal{N}_t$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\mu}_{ij}(t) - \mu_{ij} \geq \frac{\sqrt{2\epsilon\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{ij}(t)} \,\middle|\, \mathcal{N}_t\right) \\
&\leq \mathbb{P}\left(\hat{\mu}_{i,j}(t) - \mu_{i,j} \geq \frac{\mu_{i,j}(\bar{N}_{i,j}(t) - N_{i,j}(t)) + \sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{i,j}(t)}\right) \\
&\leq \epsilon e^{1-\epsilon}\log t.
\end{aligned}
$$

Thus, we obtain one direction of Eq. (5.2). Similarly, for the reversed direction, we have

$$\begin{aligned}
&\mathbb{P}\left(\mu_{ij}(t) - \bar{\mu}_{ij} \geq \frac{\sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{\bar{N}_{ij}(t)}\right) \\
&= \mathbb{P}\left(\frac{\mu_{ij}(t)\bar{N}_{i,j}(t)}{N_{i,j}(t)} - \hat{\mu}_{ij} \geq \frac{\sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{ij}(t)}\right) \\
&= \mathbb{P}\left(\mu_{ij}(t) - \hat{\mu}_{ij} \geq \frac{\mu_{i,j}(N_{i,j}(t) - \bar{N}_{i,j}(t)) + \sqrt{\frac{\epsilon}{2}\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{ij}(t)}\right) \\
&\leq \epsilon e^{1-\epsilon}\log t.
\end{aligned}$$

Replacing $\left(N_{i,j}(t) - \bar{N}_{i,j}(t)\right)$ by $\|\boldsymbol{T}_{i,j}(t)\|_1 \max_{j\in[M]} \sqrt{\frac{\epsilon_1 \ln t}{\mu_{i,j}^2 T_{i,j,k}(t)}}$ with $\epsilon_1 = \epsilon/\ln t$, conditioned on $\mathcal{N}_t$, we have

$$\mathbb{P}\left(\mu_{ij}(t) - \hat{\mu}_{ij} \geq \frac{\sqrt{2\epsilon\|\boldsymbol{T}_{i,j}(t)\|_1}}{N_{i,j}(t)} \,\middle|\, \mathcal{N}_t\right) \leq \epsilon e^{1-\epsilon}\log t. \tag{C.3}$$

□

Proof of Theorem 12

Theorem 12. *(Personalized treatment with GreedyRank)* Setting $\varepsilon_t = t^{-1/2}$, the expected regret of GreedyRank Option 1 at any time step t can be bounded as

follows:

$$\mathbb{E}[R(t)] \leq 2N\sqrt{t} + \sum_{i\in[N]} \frac{8C_\rho MK\sqrt{\zeta_i t}\ln t}{(1-1/C)\min_j \mu_{i,j}} + \mathcal{O}(1),$$

where $C, C_\rho > 1$ are problem-dependent constants.

We will leverage the following lemma and proposition in the proof.

Proposition 2 (McDiarmid's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent (not necessarily identical in distribution) random variables. Let $f : \mathcal{X}_1 \times \cdots \mathcal{X}_n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$-bounded difference property: for every $i = 1, \ldots, n$ and every $(x_1, \ldots, x_n), (x_1', \ldots, x_n') \in \mathcal{X}_1 \times \cdots \mathcal{X}_n$ that differ only in the i-th coordinate ($x_j = x_j'$ for all $j \neq i$), we have $|f(x_1, \ldots, x_n) - f(x_1', \ldots, x_n')| \leq c_i$. Then , for any $t > 0$, we have*

$$\mathbb{P}\left(|f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)]| \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma 22. *At time t, for any user $i \in [N]$ and any $\epsilon \geq 0$, the estimated user arrival rate $\hat{\zeta}_i(t)$ in Greedy Ranking satisfies the following*

$$\mathbb{P}\left(t \cdot \left|\hat{\zeta}_i(t) - \zeta_i\right| \geq \epsilon\right) \leq \exp\left(-2t\epsilon^2\right).$$

Proof. We start from evaluating the initialization phase. Since the initialization performs a round-robin sampling, then, at time $t > 1$, for each user i, arm j, position k, we have

$$\mathbb{E}[S_{i,j,k}(t) - S_{i,j,k}(t-1)] = \frac{\zeta_i \mu_{i,j}}{M}$$

By the definition of time step t_0, we have

$$\mathbb{E}[t_0] \leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}}.$$

Next, we analyze the exploration. Denote the cumulative number of exploration times that policy Greedy Ranking performs up to time t by ξ_t, then we observe the following

$$\mathbb{E}[\xi_t] = \sum_{s\leq t} \varepsilon_s \geq t \cdot \varepsilon_t.$$

Assume that there exists a constant $C > 1$ such that $\varepsilon_t \geq Ct^{-1/2}$. Then, by Hoeffding's Inequality, we obtain

$$\mathbb{P}\left(\frac{1}{t} \cdot \left|\mathbb{E}[\xi_t] - \xi_t\right| \geq \frac{\varepsilon_t}{C}\right) \leq \exp\left(-\frac{2t\varepsilon_t^2}{C^2}\right)$$

By the relations $\mathbb{E}[\xi_t] \geq t\varepsilon_t$ and $\varepsilon_t \geq Ct^{-1/2}$, we have

$$\mathbb{P}\left(t\varepsilon_t - \xi_t \geq \frac{t\varepsilon_t}{C}\right) \leq \exp\left(-\frac{2t\varepsilon_t^2}{C^2}\right) \leq \exp\left(-2t^{1/2}\right).$$

Define an event $\mathcal{F}_t$ regarding ξ_t as follows: at time t, $\xi_t \geq (1 - 1/C) \cdot t\varepsilon_t$, and we have

$$\mathbb{P}(\mathcal{F}_t^C) = \mathbb{P}\left(\xi_t \leq t\varepsilon_t - \frac{t\varepsilon_t}{C}\right) \leq \exp\left(-2t^{1/2}\right) < \exp(-2\ln t) = \frac{1}{t^2}.$$

In policy GreedyRank, we perform round robin over users and arms during exploration, thus, conditioned on event $\mathcal{F}_t$, for each user $i \in [N]$ and each arm $j \in [M]$ we

have

$$\|\boldsymbol{T}_{i,j}(t)\|_1 \geq \frac{\xi_t}{NM} \geq \frac{(1-1/C)\cdot t\varepsilon_t}{NM}.$$

The analysis of exploitation requires a high probability bound of $|\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)-\mu_{i,j}\cdot\rho_{i,\sigma_t(j)}|$. We note that given a policy $\{\sigma_t\}_t$, the expected value of $\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)$ depends on the policy, and may change over time. In other words, the expected value of $\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)$ is some function of a policy-related sequence of samples. Thus, to bound the deviation of $\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)$ to its expectation, we fix a user type i and an arm j, and we use McDiarmid's inequality as stated in Proposition 2. Define an event $\mathcal{P}_t$ regarding $\hat{\boldsymbol{\rho}}(t)$ as follows: at time t, for any user $i\in[N]$ and any position $k\in[K]$, it holds that $|\hat{\rho}_{i,k}(t)-\rho_{i,k}(t)| < \max_{j\in[M]}\sqrt{2\ln t/(\mu_{i,j}^2 T_{i,j,k}(t))}$. Define an event $\mathcal{U}_t$ regarding $\hat{\boldsymbol{\mu}}(t)$ as follows: at time t, for any user i and any arm j, it holds that $|\hat{\mu}_{i,j}(t)-\mu_{i,j}| < \sqrt{2}/N_{i,j}(t)$. Then, given any policy $\{\sigma_t\}_t$, at time t, conditioned on event $\mathcal{P}_t\cap\mathcal{U}_t$, for any user i and any arm j, the random variable $\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)$ can change by at most $\sqrt{4\ln t}/\left(N_{i,j}(t)\min_j\sqrt{\mu_{i,j}^2 T_{i,j,k}(t)}\right)$. By Proposition 2, we have

$$\mathbb{P}\left(\left|\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)-\mu_{i,j}\cdot\rho_{i,\sigma_t(j)}\right| \geq \frac{2\ln t}{\mu_{i,j}N_{i,j}(t)} \;\middle|\; \mathcal{P}_t,\mathcal{U}_t\right) \leq t^{-2},$$

where $\mathbb{P}(\mathcal{P}_t^C)\leq MKt^{-4}$, and $\mathbb{P}(\mathcal{U}_t^C)\lesssim t^{-1/2}\exp(1-t^{-1/2})\log t$. By union bound, we obtain

$$\mathbb{P}\left(\left|\sum_{j\in\mathrm{M}_\sigma}\hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t)-\sum_{j\in\mathrm{M}_\sigma}\mu_{i,j}\cdot\rho_{i,\sigma_t(j)}\right| \geq \sum_{j\in\mathrm{M}_\sigma}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)} \;\middle|\; \mathcal{P}_t,\mathcal{U}_t\right) \leq Kt^{-2}.$$

Now, denote $\Gamma_i(\sigma) = \sum_{j\in \mathrm{M}_\sigma} \mu_{i,j} \cdot \rho_{i,\sigma(j)}$, and denote $\hat{\Gamma}_i^t(\sigma) = \sum_{j\in \mathrm{M}_\sigma} \hat{\mu}_{i,j}(t) \cdot \hat{\rho}_{i,\sigma(j)}(t)$, by union bound we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j\in \mathrm{M}_\sigma} \hat{\mu}_{i,j}(t)\cdot\hat{\rho}_{i,\sigma_t(j)}(t) - \sum_{j\in \mathrm{M}_\sigma} \mu_{i,j}\cdot\rho_{i,\sigma_t(j)}\right| \geq \sum_{j\in \mathrm{M}_\sigma} \frac{2\ln t}{\mu_{i,j}N_{i,j}(t)} \;\middle|\; \mathcal{P}_t, \mathcal{U}_t\right)\\
&= \mathbb{P}\left(\left|\hat{\Gamma}_i^t(\sigma) - \Gamma_i(\sigma)\right| \geq \sum_{j\in \mathrm{M}_\sigma} \frac{2\ln t}{\mu_{i,j}N_{i,j}(t)} \;\middle|\; \mathcal{P}_t, \mathcal{U}_t\right)\\
&\leq Kt^{-2}.
\end{aligned}$$

Define an event $\mathcal{Y}_t^i$ regarding the estimated CUF $\hat{\Gamma}_i^t(\sigma_t)$ as follows: for any policy $\{\sigma_t\}_t$ and any user type i, it holds that $|\hat{\Gamma}_i(\sigma_t) - \Gamma_i(\sigma_t)| < \sum_{j\in \mathrm{M}_{\sigma_t}} 2\ln t/(\mu_{i,j}N_{i,j}(t))$. Conditioned on event $\mathcal{F}_t \cap \mathcal{Y}_t^1 \cap \ldots \cap \mathcal{Y}_t^N$, we have

$$\begin{aligned}
\Gamma_i(\sigma^*) - \Gamma_i(\sigma_t) &= \left[\Gamma_i(\sigma^*) - \hat{\Gamma}_i^t(\sigma^*)\right] + \left[\hat{\Gamma}_i^t(\sigma^*) - \hat{\Gamma}_i^t(\sigma_t)\right] + \left[\hat{\Gamma}_i^t(\sigma_t) - \Gamma_i(\sigma_t)\right]\\
&\leq \left[\Gamma_i(\sigma^*) - \hat{\Gamma}_i^t(\sigma^*)\right] + \left[\hat{\Gamma}_i^t(\sigma_t) - \Gamma_i(\sigma_t)\right]\\
&\leq \sum_{\sigma\in\{\sigma^*,\sigma_t\}} \left|\hat{\Gamma}_i^t(\sigma) - \Gamma_i(\sigma)\right|\\
&\leq \frac{4C_\rho NMK\ln t}{\min_j \mu_{i,j}(1-1/C)t\varepsilon_t},
\end{aligned}$$

where $C_\rho \geq 1$ is a constant such that $\|\boldsymbol{T}_{i,j}(t)\|_1 = C_\rho N_{i,j}(t)$, which depends on the position preference distribution and policy, independent of K, M, N. Then, we

evaluate the regret of personalized treatment GreedyRank up to time t as follows

$$
\begin{aligned}
&\mathbb{E}[R(t)] \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \mathbb{E}\left[\varepsilon_s + (1-\varepsilon_s) \sum_{i\in[N]} \left(\Gamma_i(\sigma^*) - \Gamma_i(\sigma_s)\right)\right] \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{\varepsilon_s + \sum_{i\in[N]} \mathbb{E}\left[\Gamma_i(\sigma^*) - \Gamma_i(\sigma_s)\middle| \mathcal{F}_s \cap \mathcal{Y}_s^1 \cap \ldots \cap \mathcal{Y}_s^N\right] + \right. \\
&\quad \left. \mathbb{P}\left(\mathcal{F}_s^C\right) + \mathbb{P}\left(\mathcal{Y}_s^{1C} \cup \ldots \cup \mathcal{Y}_s^{NC}\right)\right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{\varepsilon_s + \sum_{i\in[N]} \mathbb{E}\left[\Gamma_i(\sigma^*) - \Gamma_i(\sigma_s)\middle| \mathcal{F}_s \cap \mathcal{Y}_s^1 \cap \ldots \cap \mathcal{Y}_s^N\right] + \right. \\
&\quad \mathbb{P}\left(\mathcal{Y}_s^{1C} \cup \ldots \cup \mathcal{Y}_s^{NC} | \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s\right) + \mathbb{P}\left(\mathcal{A}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{P}_s^C | \mathcal{F}_s\right) + \\
&\quad \left. \mathbb{P}\left(\mathcal{U}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{F}_s^C\right)\right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{\varepsilon_s + \sum_{i\in[N]} \mathbb{E}\left[\Gamma_i(\sigma^*) - \Gamma_i(\sigma_s)\middle| \mathcal{F}_s \cap \mathcal{Y}_s^1 \cap \ldots \cap \mathcal{Y}_s^N\right] + \right. \\
&\quad N \cdot \mathbb{P}\left(\mathcal{Y}_s^{1C} | \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s\right) + \mathbb{P}\left(\mathcal{A}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{P}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{U}_s^C | \mathcal{F}_s \cap \mathcal{N}_s\right) + \\
&\quad \left. \mathbb{P}\left(\mathcal{N}_s^C\right) + \mathbb{P}\left(\mathcal{F}_s^C\right)\right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=1}^{t} \left\{\varepsilon_s + \frac{4C_\rho NMK \ln s}{\min \mu_{i,j}(1-1/C)s\varepsilon_s} + f(1)\delta_s N + N^2 K s^{-2} + s^{-2\ln s} + MKs^{-4} + \right. \\
&\quad \left. s^{-1/2}\exp(1-s^{-1/2})\ln s + MKs^{-2} + s^{-2}\right\}
\end{aligned}
$$

Setting $\varepsilon_t = Nt^{-1/2}$, we obtain the regret $\mathbb{E}[R(t)]$ of personalized GreedyRank upper bounded by

$$\mathbb{E}[R(t)] \leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}} + 2Nt^{\frac{1}{2}} + \sum_{i\in[N]} \frac{8C_\rho MK\sqrt{\zeta_i t}\ln t}{(1-1/C)\min \mu_{i,j}} + \mathcal{O}(1).$$

□

Proof of Lemma 22

Proof. At any time t, the user $i \in [N]$ arrives with probability ζ_i. Then, by Hoeffding's Inequality, for any $\epsilon \geq 0$, the estimated user arrival rate $\hat{\zeta}_i(t)$ satisfies the following

$$\mathbb{P}\left(t \cdot \left|\hat{\zeta}_i(t) - \zeta_i\right| \geq \epsilon\right) \leq \exp\left(-2t\epsilon^2\right).$$

□

Proof of Theorem 13

Theorem 13. *(Personalized treatment with UCBRank)* Setting $a_t \in (2/\min \mu_{i,j}, \sqrt{t/\ln t}]$, the expected regret of UCBRank Option 1 at any time step t can be bounded as follows:

$$\mathbb{E}[R(t)] \leq \sum_{i\in[N]} \frac{2C_\rho MK\sqrt{\zeta_i t \ln t}}{\min_j \Delta_{i,j}} + \mathcal{O}(1).$$

Proof. For user type i, define a "bad" event $\mathcal{E}_i^t$ regarding policy $\{\sigma_t\}_t$ as follows:

$$\mathcal{E}_i^t \triangleq \left\{ \hat{\Gamma}_i^t(\sigma_t) + \sum_{j \in \mathrm{M}_{\sigma_t}} \frac{a_t \ln t}{N_{i,j}(t)} \geq \hat{\Gamma}_i^t(\sigma_i^*) + \sum_{j \in \mathrm{M}_{\sigma_i^*}} \frac{a_t \ln t}{N_{i,j}(t)} \right\},$$

then event $\mathcal{E}_i^t$ is necessary and sufficient for event $\{\sigma_i^* \neq \sigma_t\}$. We observe that event $\mathcal{E}_i^t$ implies that at least one of the following events must hold:

$$\hat{\Gamma}_i^t(\sigma_t^*) \leq \Gamma_i(\sigma_i^*) - \sum_{j \in \mathrm{M}_{\sigma_i^*}} \frac{a_t \ln t}{N_{i,j}(t)} \tag{C.4}$$

$$\hat{\Gamma}_i^t(\sigma_t) \geq \Gamma_i(\sigma_t) + \sum_{j \in \mathrm{M}_{\sigma_t}} \frac{a_t \ln t}{N_{i,j}(t)} \tag{C.5}$$

$$\Gamma(\sigma_i^*) < \Gamma_i(\sigma_t) + \sum_{j \in \mathrm{M}_{\sigma_t}} \frac{2a_t \ln t}{N_{i,j}(t)}. \tag{C.6}$$

Conditioned on event $\mathcal{Y}_t^1 \cap \ldots \cap \mathcal{Y}_t^N$, we obtain that the event (C.4) and (C.5) both happen with probability zero if $a_t \geq 2/\min \mu_{i,j}$. For event (C.6), when $N_{i,j}(t) = 2K\sqrt{t \ln t}/\min_j \Delta_{i,j}$ and $a_t \leq \sqrt{t/\ln t}$, we have the following with probability one:

$$\sum_{j \in \mathrm{M}_{\sigma_t}} \frac{2a_t \ln t}{N_{i,j}(t)} = \frac{a_t \min_j \Delta_{i,j}}{\sqrt{t/\ln t}} \leq \min_j \Delta_{i,j} \leq \Gamma_i(\sigma_i^*) - \Gamma_i(\sigma_t).$$

Therefore, we obtain the regret of personalized UCBRank $\mathbb{E}[R(t)]$ upper bounded as follows:

$$
\begin{aligned}
\mathbb{E}[R(t)] &= \sum_{s=1}^{t} \sum_{i\in[N]} \mathbb{E}\left[\Gamma_i(\sigma_i^*) - \Gamma_i(\sigma_s)\middle|\mathbb{P}\left(\sigma_i^* \neq \sigma_s\right)\right] \cdot \mathbb{P}\left(\sigma_i^* \neq \sigma_s\right) \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \Bigg(\sum_{i\in[N]} \mathbb{P}\left(\sigma_i^* \neq \sigma_s \middle| \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s \cap \mathcal{Y}_s^1 \cap \ldots \cap \mathcal{Y}_s^N\right) \\
&\quad + \mathbb{P}(\mathcal{A}_s^C) + \mathbb{P}(\mathcal{P}_s^C) + \mathbb{P}(\mathcal{U}_s^C) + \mathbb{P}\left(\mathcal{Y}_s^{1C} \cup \ldots \cup \mathcal{Y}_s^{NC}\right) \Bigg) \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \Bigg(\sum_{i\in[N]} \mathbb{P}\left(\sigma^* \neq \sigma_s \middle| \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s \cap \mathcal{Y}_s^1 \cap \ldots \cap \mathcal{Y}_s^N\right) \\
&\quad + \mathbb{P}(\mathcal{A}_s^C) + \mathbb{P}(\mathcal{P}_s^C) + \mathbb{P}(\mathcal{U}_s^C | \mathcal{N}_s) + \mathbb{P}(\mathcal{N}_s^C) + \mathbb{P}\left(\mathcal{Y}_s^{1C} \cup \ldots \cup \mathcal{Y}_s^{NC}\right) \Bigg) \\
&\leq \mathbb{E}[t_0] + \sum_{i\in[N]} \frac{2C_\rho MK\sqrt{\zeta_i t \ln t}}{\min_j \Delta_{i,j}} + \sum_{s=t_1}^{t} \Big[N^2 K s^{-2} + s^{-2\ln s} + MKs^{-4} \\
&\quad + s^{-1/2}\exp(1-s^{-1/2})\ln s + MKs^{-2} \Big] \\
&\leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}} + \sum_{i\in[N]} \frac{2C_\rho MK\sqrt{\zeta_i t \ln t}}{\min_j \Delta_{i,j}} + \mathcal{O}(1).
\end{aligned}
$$

□

Proof of Theorem 14

Theorem 14. *(Equal Treatment with GreedyRank)* Setting $\varepsilon_t = Nt^{-1/2}$, with a δ_t-approximate solution to the maximization problem in GreedyRank Option 2, the

expected regret of Fair-GreedyRank at any time step t can be bounded by:

$$\mathbb{E}[R(t)] \leq 2Nt^{\frac{1}{2}} + \frac{8L_U C_\rho NMK}{(1-1/C)\min \mu_{i,j}} t^{\frac{1}{2}} \ln t + \sum_{s=1}^{t} U(1) N \delta_s + \mathcal{O}(1),$$

and for $\delta_t = \mathcal{O}(t^{-1})$, we have: $\mathbb{E}[R(t)] = \mathcal{O}\left(8L_U NMKt^{\frac{1}{2}} \log t / \min \mu_{i,j}\right)$.

Proof. Similar with the proof of Theorem 12, we obtain the regret of the initialization phase as upper bounded by

$$\mathbb{E}[t_0] \leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}},$$

We now analyze the exploitation. We first obtain a concentration of the estimated CUF $\hat{\Gamma}_t(\sigma)$. Since the utility function f is L_f-Lipschitz continuous, then for any user i and any policy $\{\sigma_t\}_t$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left| f\left(\hat{\Gamma}_i^t(\sigma_t)\right) - f\left(\Gamma_i(\sigma_t)\right) \right| \geq L_f \sum_{j\in \mathrm{M}_{\sigma_t}} \frac{2\ln t}{\mu_{i,j} N_{i,j}(t)} \;\middle|\; \mathcal{P}_t, \mathcal{U}_t \right) \\
&\leq \mathbb{P}\left(L_f \cdot \left| \hat{\Gamma}_i^t(\sigma_t) - \Gamma_i(\sigma_t) \right| \geq L_f \cdot \sum_{j\in \mathrm{M}_{\sigma_t}} \frac{2\ln t}{\mu_{i,j} N_{i,j}(t)} \;\middle|\; \mathcal{P}_t, \mathcal{U}_t \right) \\
&\leq Kt^{-2}. \qquad \text{(C.7)}
\end{aligned}$$

Define an event $\mathcal{A}_t$ regarding estimated user arrival rate $\hat{\boldsymbol{\zeta}}_t$ as follows: at time t, for any user $i \in [N]$, there exists $\epsilon_1 \geq 0$ such that $|\hat{\zeta}_i(t) - \zeta_i| < \epsilon_1$. Then, conditioned on

event $\mathcal{A}_t \cap \mathcal{P}_t \cap \mathcal{U}_t$, for user $i \in [N]$, we have

$$\begin{aligned}
&\mathbb{P}\left(\hat{\zeta}_i(t)\cdot\left|f\left(\hat{\Gamma}_i^t(\sigma_t)\right)-f\left(\Gamma_i(\sigma_t)\right)\right|\geq\hat{\zeta}_i(t)L_f\cdot\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\;\middle|\;\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right)\\
\geq&\,\mathbb{P}\left(\hat{\zeta}_i(t)\cdot\left|f\left(\hat{\Gamma}_i^t(\sigma_t)\right)-f\left(\Gamma_i(\sigma_t)\right)\right|\geq L_f\cdot\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\;\middle|\;\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right)\\
\geq&\,\mathbb{P}\left(\left|\hat{\zeta}_i(t)f\left(\hat{\Gamma}_i^t(\sigma_t)\right)-(\zeta_i+\epsilon_2)f\left(\Gamma_i(\sigma_t)\right)\right|\geq L_f\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\;\middle|\;\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right)\\
\geq&\,\mathbb{P}\left(\left|\hat{\zeta}_i(t)f\left(\hat{\Gamma}_i^t(\sigma_t)\right)-\zeta_i f\left(\Gamma_i(\sigma_t)\right)\right|\geq\epsilon_2 f(1)+L_f\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\middle|\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right),
\end{aligned}\tag{C.8}$$

and by Eq. (C.7), we have (C.8) $\leq Kt^{-2}$. By union bound, we obtain

$$\mathbb{P}\left(\left|\hat{\Gamma}_t(\sigma_t)-\Gamma(\sigma_t)\right|\geq\sum_{i\in[N]}\epsilon_1 f(1)+L_f\sum_{i\in[N]}\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\;\middle|\;\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right)\leq NKt^{-2}.\tag{C.9}$$

Specifically, setting $\epsilon_1=\sum_{j\in\mathrm{M}_{\sigma_t}}2L_f\ln t/\left(f(1)\mu_{i,j}N_{i,j}(t)\right)$, we obtain

$$\mathbb{P}\left(\left|\hat{\Gamma}_t(\sigma_t)-\Gamma(\sigma_t)\right|\geq 2L_f\sum_{i\in[N]}\sum_{j\in\mathrm{M}_{\sigma_t}}\frac{2\ln t}{\mu_{i,j}N_{i,j}(t)}\;\middle|\;\mathcal{A}_t,\mathcal{P}_t,\mathcal{U}_t\right)\leq NKt^{-2},\tag{C.10}$$

in which case, we have the probability of event $\mathbb{P}(\mathcal{A}_t^C)\leq t^{-2C_1\ln t}$ with constant $C_1>0$.

We proved that in policy GreedyRank, conditioned on event $\mathcal{F}_t$, for user i and arm j, we have

$$\mathbb{P}\left(t\varepsilon_t - \xi_t \geq \frac{t\varepsilon_t}{C}\right) \leq \exp\left(-\frac{2t\varepsilon_t^2}{C^2}\right) \leq \exp\left(-2t^{1/2}\right).$$

Then, define an event $\mathcal{Y}_t$ regarding estimated CUF $\hat{\Gamma}_t(\sigma)$ as follows: for any policy $\{\sigma_t\}_t$, it holds that $|\hat{\Gamma}_t(\sigma) - \Gamma(\sigma_t)| < 2L_f \sum_{i\in[N]} \sum_{j\in \mathrm{M}_{\sigma_t}} 2\ln t/(\mu_{i,j} N_{i,j}(t))$. Conditioned on event $\mathcal{F}_t \cap \mathcal{Y}_t$, we have

$$\Gamma(\sigma^*) - \Gamma(\sigma_t) = \left[\Gamma(\sigma^*) - \hat{\Gamma}_t(\sigma^*)\right] + \left[\hat{\Gamma}_t(\sigma^*) - \hat{\Gamma}_t(\sigma_t)\right] + \left[\hat{\Gamma}_t(\sigma_t) - \Gamma(\sigma_t)\right].$$

Specifically, $\hat{\Gamma}_t(\sigma_t)$ is defined as being maximized by permutation σ_t. If we use an approximate solution with a factor δ_t of being optimal, then we have

$$\begin{aligned}
\Gamma(\sigma^*) - \Gamma(\sigma_t) &\leq \left[\Gamma(\sigma^*) - \hat{\Gamma}_t(\sigma^*)\right] + \left[\hat{\Gamma}_t(\sigma^*) - (1-\delta_t)\hat{\Gamma}_t(\sigma_t)\right] + \left[\hat{\Gamma}_t(\sigma_t) - \Gamma(\sigma_t)\right] \\
&\leq f(1)\delta_t N + \sum_{\sigma\in\{\sigma^*,\sigma_t\}} \left|\Gamma(\sigma) - \hat{\Gamma}_t(\sigma)\right| \\
&\leq f(1)\delta_t N + \frac{4L_f C_\rho N^2 M K \ln t}{\min \mu_{i,j}(1-1/C) t\varepsilon_t}.
\end{aligned}$$

Then, we evaluate the regret of equal treatment GreedyRank up to time t as follows

$$
\begin{aligned}
\mathbb{E}[R(t)] &\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \mathbb{E}\left[\varepsilon_s + (1-\varepsilon_s)\big(\Gamma(\sigma^*) - \Gamma(\sigma_s)\big)\right] \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{ \varepsilon_s + \mathbb{E}\left[\Gamma(\sigma^*) - \Gamma(\sigma_s) \middle| \mathcal{F}_s \cap \mathcal{Y}_s\right] + \mathbb{P}\left(\mathcal{F}_s^C\right) + \mathbb{P}\left(\mathcal{Y}_s^C\right) \right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{ \varepsilon_s + \mathbb{E}\left[\Gamma(\sigma^*) - \Gamma(\sigma_s) \middle| \mathcal{F}_s \cap \mathcal{Y}_s\right] + \mathbb{P}\left(\mathcal{Y}_s^C | \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s\right) \right. \\
&\quad \left. + \mathbb{P}\left(\mathcal{A}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{P}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{U}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{F}_s^C\right) \right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \left\{ \varepsilon_s + \mathbb{E}\left[\Gamma(\sigma^*) - \Gamma(\sigma_s) \middle| \mathcal{F}_s \cap \mathcal{Y}_s\right] + \mathbb{P}\left(\mathcal{Y}_s^C | \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s\right) \right. \\
&\quad \left. + \mathbb{P}\left(\mathcal{A}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{P}_s^C | \mathcal{F}_s\right) + \mathbb{P}\left(\mathcal{U}_s^C | \mathcal{F}_s \cap \mathcal{N}_s\right) + \mathbb{P}\left(\mathcal{N}_s^C\right) + \mathbb{P}\left(\mathcal{F}_s^C\right) \right\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=1}^{t} \left\{ \varepsilon_s + \frac{4L_f C_\rho N^2 MK \ln s}{\min \mu_{i,j} (1-1/C) s \varepsilon_s} + f(1)\delta_s N + NKs^{-2} + s^{-2\ln s} \right. \\
&\quad \left. + MKs^{-4} + s^{-1/2}\exp(1-s^{-1/2})\ln s + MKs^{-2} + s^{-2} \right\}
\end{aligned}
$$

Setting $\varepsilon_t = \Theta(N \cdot t^{-1/2})$, we obtain the regret $\mathbb{E}[R(t)]$ of policy GreedyRank upper bounded by

$$
\mathbb{E}[R(t)] \leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}} + 2Nt^{\frac{1}{2}} + \frac{8L_f C_\rho NMK}{(1-1/C)\min \mu_{i,j}} t^{\frac{1}{2}} \ln t + f(1)\delta_t N t + \mathcal{O}(1).
$$

□

Proof of Theorem 15

Theorem 15. *(Equal Treatment with UCBRank)* With any δ_t-approximate solution to the maximization problem in UCBRank Option 2, setting $\delta = \mathcal{O}(\sqrt{\log t/t})$ and $a_t \in (2L_U/\min \mu_{i,j}, \sqrt{t/\ln t}]$, the expected regret of UCBRank at any time step t can be bounded as follows:

$$\mathbb{E}[R(t)] = \mathcal{O}\left(\frac{N^2 MK\sqrt{t\log t}}{\Delta_\Gamma}\right).$$

Proof. Define a "bad" event $\mathcal{E}_t$ regarding policy $\{\sigma_t\}_t$ as follows:

$$\mathcal{E}_t \triangleq \left\{\hat{\Gamma}_t(\sigma_t) + \sum_{i\in[N]}\sum_{j\in \mathrm{M}_{\sigma_t}} \frac{a_t \ln t}{N_{i,j}(t)} \geq \hat{\Gamma}_t(\sigma^*) + \sum_{i\in[N]}\sum_{j\in \mathrm{M}_{\sigma^*}} \frac{a_t \ln t}{N_{i,j}(t)}\right\},$$

then event $\mathcal{E}_t$ is necessary and sufficient for event $\{\sigma^* \neq \sigma_t\}$. We observe that event $\mathcal{E}_t$ implies that at least one of the following events must hold:

$$\hat{\Gamma}_t(\sigma^*) \leq \Gamma(\sigma^*) - \sum_{i\in[N]}\sum_{j\in \mathrm{M}_{\sigma^*}} \frac{a_t \ln t}{N_{i,j}(t)} \tag{C.11}$$

$$\hat{\Gamma}_t(\sigma_t) \geq \Gamma(\sigma_t) + \sum_{i\in[N]}\sum_{j\in \mathrm{M}_{\sigma_t}} \frac{a_t \ln t}{N_{i,j}(t)} \tag{C.12}$$

$$\Gamma(\sigma^*) < \Gamma(\sigma_t) + \sum_{i\in[N]}\sum_{j\in \mathrm{M}_{\sigma_t}} \frac{2a_t \ln t}{N_{i,j}(t)}. \tag{C.13}$$

Conditioned on event $\mathcal{Y}_t$, we obtain that the event (C.11) and (C.12) both happen with probability zero if $a_t \geq 4L_f/\min \mu_{i,j}$ and the following condition is satisfied:

$$\delta_t \leq \frac{a_t - 4L_f/\min \mu_{i,j}}{f(1)N} \cdot \sum_{i\in[N]} \sum_{j\in \mathrm{M}_{\sigma_t}} \frac{\ln t}{N_{i,j}(t)}. \tag{C.14}$$

For event (C.13), when $N_{i,j}(t) = 2NK\sqrt{t \ln t}/\Delta_\Gamma$ and $a_t \leq \sqrt{t/\ln t}$, we have the following with probability one:

$$\sum_{i\in[N]} \sum_{j\in \mathrm{M}_{\sigma_t}} \frac{2a_t \ln t}{N_{i,j}(t)} = \frac{a_t \Delta_\Gamma}{\sqrt{t/\ln t}} \leq \Delta_\Gamma \leq \Gamma(\sigma^*) - \Gamma(\sigma_t).$$

Therefore, if $\|\boldsymbol{T}_{i,j}(t)\|_1 \geq 2C_\rho NK\sqrt{t \ln t}/\Delta_\Gamma$, $a_t \in (2L_f/\min \mu_{i,j}, \sqrt{t/\ln t}]$ and condition (C.14) is satisfied, we obtain the regret of equal treatment UCBRank

$\mathbb{E}[R(t)]$ upper bounded as follows:

$$\begin{aligned}
\mathbb{E}[R(t)] &= \sum_{s=1}^{t} \mathbb{E}\left[\Gamma(\sigma^*) - \Gamma(\sigma_s)\middle|\mathbb{P}\left(\sigma^* \neq \sigma_s\right)\right] \cdot \mathbb{P}\left(\sigma^* \neq \sigma_s\right) \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \Bigg\{ f(1)N \cdot \Bigg(\mathbb{P}\left(\sigma^* \neq \sigma_s \middle| \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s \cap \mathcal{Y}_s\right) \\
&\quad + \mathbb{P}(\mathcal{A}_s^C) + \mathbb{P}(\mathcal{P}_s^C) + \mathbb{P}(\mathcal{U}_s^C) + \mathbb{P}(\mathcal{Y}_s^C) \Bigg) \Bigg\} \\
&\leq \mathbb{E}[t_0] + \sum_{s=t_0}^{t} \Bigg\{ f(1)N \cdot \Bigg(\mathbb{P}\left(\sigma^* \neq \sigma_s \middle| \mathcal{A}_s \cap \mathcal{P}_s \cap \mathcal{U}_s\right) \\
&\quad + \mathbb{P}(\mathcal{A}_s^C) + \mathbb{P}(\mathcal{P}_s^C) + \mathbb{P}(\mathcal{U}_s^C | \mathcal{N}_s) + \mathbb{P}(\mathcal{N}_s^C) + \mathbb{P}(\mathcal{Y}_s^C) \Bigg) \Bigg\} \\
&\leq \mathbb{E}[t_0] + \frac{2C_\rho N^2 MK\sqrt{t \ln t}}{\Delta_\Gamma} + \sum_{s=t_1}^{t} \Bigg\{ f(1)N \Bigg(NKs^{-2} + s^{-2\ln s} \\
&\quad + MKs^{-4} + s^{-1/2}\exp(1 - s^{-1/2})\ln s + MKs^{-2} \Bigg) \Bigg\} \\
&\leq \sum_{i\in[N]} \sum_{j\in[M]} \frac{MK}{\zeta_i \mu_{i,j}} + \frac{2C_\rho N^2 MK\sqrt{t \ln t}}{\Delta_\Gamma} + C_1 f(1) N \ln t + \mathcal{O}(1).
\end{aligned}$$

□