Incentivizing Exploration with Unbiased Histories

author names withheld

Editor: Under Review for COLT 2019

Abstract

In a social learning setting, there is a set of actions, each of which has a payoff that depends on a hidden state of the world. A sequence of agents each chooses an action with the goal of maximizing payoff given estimates of the state of the world. A disclosure policy tries to coordinate the choices of the agents by sending messages about the history of past actions. The goal of the algorithm is to minimize the regret of the action sequence.

In this paper, we study a particular class of disclosure policies that use messages, called unbiased subhistories, consisting of the actions and rewards from a subsequence of past agents, where the subsequence is chosen ahead of time. One trivial message of this form contains the full history; a disclosure policy that chooses to use such messages risks inducing herding behavior among the agents and thus has regret linear in the number of rounds. Our main result is a disclosure policy using unbiased subhistories that obtains regret $\tilde{O}(\sqrt{T})$. We also exhibit simpler policies with higher, but still sublinear, regret. These policies can be interpreted as dividing a sublinear number of agents into constant-sized focus groups, whose histories are then fed to future agents.

Keywords: incentivizing exploration, multi-armed bandit, social learning

1. Introduction

In the classic literature on multi-armed bandits, an agent repeatedly selects one of a set of actions, each of which has a payoff drawn from an unknown fixed distribution. Over time, she can trade off exploitation, in which she picks an action to maximize her expected reward, with exploration, in which she takes potentially sub-optimal actions to learn more about their rewards. By coordinating her actions across time, she can guarantee an average reward which converges to that of the optimal action in hindsight at a rate proportional to the inverse square-root of the time horizon.

In many decision problems of interest, the actions are not chosen by a single agent, as above, but rather a sequence of agents. This is particularly common in social learning settings such as online websites, where a population of users try to learn about the content of the site. In such settings, each agent will choose an exploitive action as the benefits of explorative actions are only accrued by future agents. For example, in online retail, products are purchased by a sequence of customers, each of which buys what she estimates to be the best available product. This behavior can cause herding, in which all agents eventually take a sub-optimal action of maximum expected payoff given the available information.

This situation can be circumvented by a centralized algorithm that induces agents to take explorative actions, an idea called incentivizing exploration. Such algorithms are often encountered in the form of recommendations and are quite common in practice. Many online websites, like Amazon, Reddit, Yelp, and Tripadvisor, among many others, use recommendation policies of some sort to help users navigate their offerings. One way recommendation policies induce exploration is to introduce payments, e.g., (Frazier et al., 2014; Kannan et al., 2017; Chen et al., 2018). For

© 2019.
example, the recommendation system of an online retailer might offer coupons to agents for trying certain products. When payments are financially or technologically infeasible, another alternative is to rely on information asymmetry, *e.g.*, (Kremer et al., 2014; Che and Hörner, 2018; Mansour et al., 2015; Bimpikis et al., 2018). Here the idea is that the centralized algorithm, often called a disclosure policy, can choose to selectively release information about the past actions and rewards to the agents in the form of a message. For example, the recommendation system of an online retailer might disclose past reviews or product rankings to the agents. Importantly, agents can not directly observe the past, but only learn about it through this message. The agent then chooses an action, using the content of the message as input.

**Our scope.** Prior work on incentivizing exploration, with or without monetary incentives, achieves much progress (more on this in “related work”), but relies heavily on the standard assumptions of Bayesian rationality and the “power to commit” (*i.e.*, users trust that the principal actually implements the policy that it claims to implement). However, these assumptions appear quite problematic in the context of recommendation systems of actual online websites such as those mentioned above. In particular, much of the prior work suggests disclosure policies that merely recommend an action to each agent, without any other supporting information, and moreover recommend exploratory actions to some randomly selected users. This works out extremely well in theory, but it is very unclear whether users would even know some complicated policy of the principal, let alone trust the principal to implement the stated policy. Even if they do know the policy and trust that it was implemented as claimed, it’s unclear whether users would react to it rationally. Several issues are in play: to wit, whether the principal intentionally uses a different disclosure policy than the claimed one (*e.g.*, because its incentives are not quite aligned with the users’), whether the principal correctly implements the policy that it wants to implement, whether the users trust the principal to make correct inferences on their behalf, and whether they find it acceptable that they may be singled out for exploration. Furthermore, regardless of how the users react to such disclosure policies, they may prefer not to be subjected to them, and leave the system.

We strive to design disclosure policies which mitigate these issues and (still) incentivize a good balance between exploration and exploitation. While some assumptions on human behavior are unavoidable, we are looking for a class of disclosure policies for which we can make plausible behavioral assumptions. Then we arrive at a concrete mathematical problem: design policies from this class so as to optimize performance, *i.e.*, the induced explore-exploit tradeoff. Our goal in terms of performance is to approach the performance of the social planner.

**Our model.** For the sake of intuition, let us revisit the full-disclosure policy that reveals the full history of observations from the previous users. We interpret it as the “gold standard”: we posit that users would trust such policy, even if they cannot verify it. Unfortunately, the full-disclosure policy is not good for our purposes, essentially because we expect users to exploit rather than explore. However, what if a disclosure policy reveals the outcomes for every tenth agent, rather than the outcomes for all agents? We posit that users would trust such policy, too. Given a large volume of data, we posit that users would not be too unhappy with having access to only a fraction this data. A crucial aspect of our intuition here is that the “subhistory” revealed to a given user comes from a subset of previous users that is chosen in advance, without looking at what happens during the execution. In particular, the subhistory is not “biased”, in the sense that the disclosure policy cannot subsample the observations in favor of a particular action.
With this intuition in mind, we define the class of unbiased-subhistory policies: disclosure policies that reveal, to each arriving agent \( t \), a subhistory consisting of the outcomes for a subset \( S_t \) of previous agents, where \( S_t \) is chosen ahead of time. Further, we impose a transitivity property: if \( t' \in S_t \), for some previous agent \( t' \), then \( S_{t'} \subset S_t \). So, agent \( t \) has all information that agent \( t' \) had at the time she chose her action. In particular, agent \( t \) does not need to second-guess which message has caused agent \( t' \) to make choose that action.

Following much of the prior work on incentivizing exploration, we do not attempt to model heterogenous agent preferences and non-stationarity. Formally, we assume that the expected reward of taking a given action \( a \), denoted \( \mu_a \), is the same for all agents, and does not change over time. Then the crucial parameter of interest, for a given action \( a \), are the number of samples \( N_a \) and the empirical mean reward \( \bar{\mu}_a \) in the observed subhistory. We consider a flexible model of agent response: for each action \( a \) an agent forms an estimate \( \hat{\mu}_a \) of the mean reward \( \mu_a \), roughly following \( \bar{\mu}_a \) but taking into account the uncertainty due to a small number of samples, and chooses an action with a largest reward estimate. We allow the reward estimates to be arbitrary otherwise, and not known to the principal.

**Regret.** We measure the performance of a disclosure policy in terms of regret, a standard notion from the literature on multi-armed bandits. Regret is defined as the difference in the total expected reward between the best fixed action and actions induced by the policy. Regret is typically studied as a function of the time horizon \( T \), which in our model is the number of agents. For multi-armed bandits, \( o(T) \) regret bounds are deemed non-trivial, and \( O(\sqrt{T}) \) regret bounds are optimal in the worst case. Regret bounds that depend on a particular problem instance are also considered. A crucial parameter then is the gap \( \Delta \), the difference between the best and second best expected reward. One can achieve \( O(\frac{1}{\Delta} \log T) \) regret rate, without knowing the \( \Delta \).

**Our results and discussion.** Our main result is a transitive, unbiased-subhistory policy which attains near-optimal \( \tilde{O}(\sqrt{T}) \) regret rate for a constant number of actions. This policy also obtains the optimal instance-dependent regret rate \( \tilde{O}(\frac{1}{\Delta}) \) for problem instances with gap \( \Delta \), without knowing the \( \Delta \) in advance. In particular, we match the regret rate achieved for incentivizing exploration with unrestricted disclosure policies (Mansour et al., 2018a).

The main challenge is that the agents still follow exploitation-only behavior, as for the full-disclosure policy, albeit based only on a portion of history. A disclosure policy controls the flow of information (who sees what), but not the content of that information.

The first step is to obtain any substantial improvement over the full-disclosure policy. We accomplish this with a relatively simple policy which runs the full-disclosure policy “in parallel” on several disjoint subsets of agents, collects all data from these runs and discloses it to all remaining agents. In practice, these subsets may correspond to multiple “focus groups”. While any single run of the full-disclosure policy may get stuck on a suboptimal arm, having these parallel runs ensure that sufficiently many of them will “get lucky” and provide some exploration. This simple policy achieves \( \tilde{O}(T^{2/3}) \) regret. Conceptually, it implements a basic bandit algorithm that explores uniformly for a pre-set number of rounds, then picks one arm for exploitation and stays with it for the remaining rounds. We think of this policy as having two “levels”: Level 1 contains the parallel runs, and Level 2 is everything else.

The next step is to implement adaptive exploration, where the exploration schedule is adapted to previous observations. This is needed to improve over the \( \tilde{O}(T^{2/3}) \) regret. As a proof of concept, we focus on the case of two actions, and upgrade the simple two-level policy with a middle level.
The agents in this new level receive the data collected in some (but not all) runs from the first level. What happens is that these agents explore only if the gap $\Delta$ between the best and second-best arm is sufficiently small, and exploit otherwise. When $\Delta$ is small, the runs in the first level do not have sufficient time to distinguish the two arms before herding on one of them. However, for each of these arms, there is some chance that it has an empirical mean reward significantly above its actual mean while the other arm has empirical mean reward significantly below its actual mean in any given first-level run. The middle-level agents observing such runs will be induced to further explore that arm, collecting enough samples for the third-level agents to distinguish the two arms. The main result extends this construction to multiple levels, connected in fairly intricate ways, obtaining optimal regret of $O(T^{1/2})$.

**Related work.** The problem of incentivizing exploration via information asymmetry was introduced in Kremer et al. (2014), under the Bayesian rationality and the (implicit) power-to-commit assumptions. The original problem – essentially, a version of our model with unrestricted disclosure policies – was largely resolved in Kremer et al. (2014) and the subsequent work (Mansour et al., 2015, 2016). Several extensions were considered: to contextual bandits (Mansour et al., 2015), to repeated games (Mansour et al., 2016), and to social networks (Bahar et al., 2016).

Several other papers study related, but technically different models: same model with time-discounted utilities (Bimpikis et al., 2018); a version with monetary incentives (Frazier et al., 2014) and moreover with heterogenous agents (Chen et al., 2018); a version with a continuous information flow and a continuum of agents (Che and Hörner, 2018); coordination of costly exploration decisions when they are separate from the “payoff-generating” decisions (Kleinberg et al., 2016; Liang and Mu, 2018; Liang et al., 2018). Scenarios with long-lived, exploring agents and no principal to coordinate them have been studied in (Bolton and Harris, 1999; Keller et al., 2005).

Full-disclosure policy, and closely related “greedy” (exploitation-only) algorithm in multi-armed bandits, have been a subject of a recent line of work (Schmit and Riquelme, 2018; Kannan et al., 2018; Bastani et al., 2018; Raghavan et al., 2018). A common theme is that the greedy algorithm performs well in theory, under substantial assumptions on heterogeneity of the agents. Yet, it suffers $\Omega(T)$ regret in the worst case.\footnote{1} Baransi et al. (2014) use the “greedy” algorithm on sub-sampled subhistories to achieve optimal regret rate for two arms. However, their result is insufficient for our motivation, as the subhistories are not unbiased or transitive, in the sense defined earlier. In particular, the sub-sampling scheme depends on the history, which may distort agents’ incentives.

Exploration-exploitation tradeoff received much attention over the past decades, usually under the rubric of “multi-armed bandits”; see (Bubeck and Cesa-Bianchi, 2012) for background. Exploration-exploitation problems with incentives issues arise in other scenarios, including dynamic pricing (Kleinberg and Leighton, 2003; Besbes and Zeevi, 2009; Badanidiyuru et al., 2018), dynamic auctions (Athey and Segal, 2013; Bergemann and Välimäki, 2010; Kakade et al., 2013), pay-per-click ad auctions (Babaioff et al., 2014; Devanur and Kakade, 2009; Babaioff et al., 2015), and human computation (Ho et al., 2016; Ghosh and Hummel, 2013; Singla and Krause, 2013).

### 2. Model and Preliminaries

We study the multi-armed bandit problem in a social learning context, in which a principal faces a sequence of $T$ myopic agents. There is a set $\mathcal{A}$ of $K$ possible actions, a.k.a. *arms*. At each round\footnote{1. This is a well-known folklore result; e.g., see (Mansour et al., 2018b) for a concrete example.}
\( t \in [T] \), a new agent \( t \) arrives, receives a message \( m_t \) from the principal, chooses an arm \( a_t \in A \), and collects a reward \( r_t \in \{0, 1\} \) that is immediately observed by the principal. The reward from pulling an arm \( a \in A \) is drawn independently from Bernoulli distribution \( D_a \) with an unknown mean \( \mu_a \). An agent does not observe anything from the previous rounds, other than the message \( m_t \). The problem instance is defined by (known) parameters \( K, T \) and the (unknown) tuple of mean rewards, \((\mu_a : a \in A)\). We are interested in regret, defined as \( \text{Reg}(T) = T \max_{a \in A} \mu_a - \sum_{t \in [T]} E[\mu_{a_t}] \). (The expectation is over the chosen arms \( a_t \), which depend on randomness in rewards, and possibly in the algorithm.) The principal chooses messages \( m_t \) according to an online algorithm called disclosure policy, with a goal to minimize regret. We assume that mean rewards are bounded away from 0 and 1, to ensure sufficient entropy in rewards. For concreteness, we posit \( \mu_a \in [\frac{1}{3}, \frac{2}{3}] \).

**Unbiased subhistories.** The subhistory for a subset of rounds \( S \subset [T] \) is defined as

\[
\mathcal{H}_S = \{ (s, a_s, r_s) : s \in S \}.
\]

\( \mathcal{H}_{[t-1]} \) is called the full history at time \( t \). The outcome for agent \( t \) is the tuple \((t, a_t, r_t)\).

We focus on disclosure policies of a particular form: the message in each round \( t \) is \( m_t = \mathcal{H}_{S_t} \) for some subset \( S_t \subset [t - 1] \). We assume that \( S_t \) is chosen ahead of time, before round 1 (and therefore does not depend on the observations \( \mathcal{H}_{t-1} \)). Such message is called unbiased subhistory, and the resulting disclosure policy is called an unbiased-history policy.

Further, we focus on disclosure policies that are transitive, in the following sense:

\[
t \in S_{t'} \Rightarrow S_t \subset S_{t'} \quad \text{for all rounds } t, t' \in [T].
\]

In words, if agent \( t' \) observes the outcome for some previous agent \( t \), then she observes the entire message revealed to that agent. In particular, agent \( t' \) does not need to second-guess which message has caused agent \( t \) to choose action \( a_t \).

A transitive unbiased-history policy can be represented as an undirected graph, where nodes correspond to rounds, and any two rounds \( t < t' \) are connected if and only if \( t \in S_{t'} \) and there is no intermediate round \( t'' \) with \( t \in S_{t''} \) and \( t'' \in S_{t'} \). This graph is henceforth called the information flow graph of the policy, or info-graph for short. We assume that this graph is common knowledge.

**Agents’ behavior.** Let us define agents’ behavior in response to an unbiased-history policy. We posit that each agent \( t \) uses its observed subhistory \( m_t \) to form a reward estimate \( \hat{\mu}_{t,a} \in [0, 1] \) for each arm \( a \in A \), and chooses an arm with a maximal estimator. (Ties are broken according to an arbitrary rule that is the same for all agents.) The basic model is that \( \hat{\mu}_{t,a} \) is the sample average for arm \( a \) over the subhistory \( m_t \), as long as it includes at least one sample for \( a \); else, \( \hat{\mu}_{t,a} \geq \frac{1}{3} \).

We allow a much more permissive model which allows agents to form arbitrary reward estimates as long as they lie within some “confidence range” of the sample average. The model is characterized by the following assumptions (made without further notice).

**Assumption 1** Reward estimates are close to empirical averages. Let \( N_{t,a} \) and \( \hat{\mu}_{t,a} \) denote the number of pulls and the empirical mean reward of arm \( a \) in subhistory \( m_t \). Then for some absolute constant \( N_{\text{est}} \in N \) and \( C_{\text{est}} = \frac{1}{16} \), and for all agents \( t \in [T] \) and arms \( a \in A \) we have

\[
\text{if } N_{t,a} \geq N_{\text{est}} \quad \text{then } \quad \left| \hat{\mu}_{t,a} - \bar{\mu}_{a} \right| < C_{\text{est}} / \sqrt{N_{t,a}}.
\]

Also, \( \bar{\mu}_{a} \geq \frac{1}{3} \) if \( N_{t,a} = 0 \). (NB: we make no assumption if \( 1 < N_{t,a} < N_{\text{est}} \).)
**Assumption 2** In each round \( t \), the estimates \( \hat{\mu}_{t,a} \) depend only on the multiset \( \{ (a_s, r_s) : s \in S_t \} \), called anonymized subhistory. Each agent \( t \) forms its estimates according to an estimate function \( f_t \) from anonymized subhistories to \([0,1]^K\), so that the estimate vector \( (\hat{\mu}_{t,a} : a \in A) \) equals \( f_t(m_t) \). This function is drawn from some fixed distribution over estimate functions.

**Connection to multi-armed bandits.** The special case when each message \( m_t \) is an arm, and the \( t \)-th agent always chooses this arm, corresponds to a standard multi-armed bandit problem with IID rewards. Thus, regret in our problem can be directly compared to regret in the bandit problem with the same mean rewards \( (\mu_a : a \in A) \). Following the literature on bandits, we define the gap parameter \( \Delta \) as the difference between the largest and second largest mean rewards.\(^2\) The gap parameter is not known to the principal (in our problem), or to the algorithm (in the bandit problem). Optimal regret rates for bandits with IID rewards are as follows (Auer et al., 2002a,b):

\[
\text{Reg}(T) \leq O \left( \min \left( \sqrt{KT \log T}, \frac{1}{K} \log T \right) \right). \tag{1}
\]

This regret bound can only be achieved using adaptive exploration: i.e., when exploration schedule is adapted to the observations. A simple example of non-adaptive exploration is the explore-then-exploit algorithm which samples arms uniformly at random for the first \( N \) rounds, for some pre-set \( N \), then chooses one arm and sticks with it till the end. Any such algorithm suffers from \( \Omega(T^{2/3}) \) regret in the worst case.\(^3\)

**Preliminaries.** We assume that \( K \) is constant, and focus on the dependence on \( T \). However, we explicitly state the dependence on \( K \), e.g., using the \( O_K(\cdot) \) notation.

Throughout the paper, we use the standard concentration and anti-concentration inequalities: resp., Chernoff Bounds and Berry-Esseen Theorem (see Appendix A). The former states that that a sum of independent random variables converges to its expectation quickly. The latter states that the CDF of an appropriately scaled average of IID random variables converges to the CDF of the standard normal distribution pointwise. In particular, the average strays far enough from its expectation with some guaranteed probability.

We use the notion of reward tape to simplify applications of (anti-)concentration inequalities. This is a \( K \times T \) random matrix with rows and columns corresponding to arms and rounds. For each arm \( a \) and round \( t \), the value in cell \( (a,t) \) is drawn independently from Bernoulli distribution \( D_a \). W.l.o.g., rewards in our model are defined by the rewards tape: the reward for the \( j \)-th pull of arm \( a \) is taken from cell \( (a,j) \) of the reward tape.

We use \( O_K(\cdot) \) notation to hide the dependence on parameter \( K \), and \( \tilde{O}(\cdot) \) notation to hide polylogarithmic factors. We denote \([T] = \{1,2,\ldots,T\}\).

### 3. Warm-up: full-disclosure paths

We first consider a disclosure policy that reveals the full history in each round \( t \), i.e., \( m_t = \mathcal{H}_{t-1} \); we call it the full-disclosure policy. The info-path for this policy is a simple path. We use this policy as a “gadget” in our constructions. Hence, we formulate it more generally:

\(^2\) Formally, the second-largest mean reward is \( \max_{a \in A, \mu(a) < \mu^*} \mu(a) \), where \( \mu^* = \max_{a \in A} \mu(a) \).

\(^3\) More generally, this lower bound applies to exploration-separating algorithms, where in each round \( t \), either the choice of an arm does not depend on the observations so far, or the reward collected in this round is not used in the subsequent rounds. The first explicit reference we know of is Babaioff et al. (2014); Devanur and Kakade (2009), but this fact has been known in the community for much longer.
Incentivizing Exploration with Unbiased Histories

**Definition 3** A subset of rounds \( S \subseteq [T] \) is called a full-disclosure path in the info-graph \( G \) if the induced subgraph \( G_S \) is a simple path, and it connects to the rest of the graph only through the terminal node \( \max(S) \), if at all.

We prove that for a constant number of arms, with constant probability, a full-disclosure path of constant length suffices to sample each arm once. We use this fact throughout.

**Lemma 4** There exist numbers \( L_F^{\text{FDP}} > 0 \) and \( p_F^{\text{FDP}} > 0 \) that depend only on \( K \), the number of arms, with the following property. Consider an arbitrary disclosure policy, and let \( S \subseteq [T] \) be a full-disclosure path in its info-graph, of length \( |S| \geq L_F^{\text{FDP}} \). Under Assumption 1, with probability at least \( p_F^{\text{FDP}} \), subhistory \( H_S \) contains at least once sample of each arm \( a \).

We provide a simple disclosure policy based on full-disclosure paths. The policy follows the “explore-then-exploit” paradigm. The “exploration phase” comprises the first \( T_1 \cdot L_F^{\text{FDP}} \) rounds, and consists of \( T_1 \) full-disclosure paths of length \( L_F^{\text{FDP}} \) each, where \( T_1 \) is a parameter. In the “exploitation phase”, each agent \( t > N \) receives the full subhistory from exploration, i.e., \( m_t = H_{[N]} \).

The info-graph for this disclosure policy is in Figure 1.

The info-graph has two “levels”, corresponding to exploration and exploitation. Accordingly, we call this policy the **two-level policy**. We show that it incentivizes the agents to perform non-adaptive exploration, and achieves a regret rate of \( \tilde{O}_K(T^{2/3}) \). The key idea is that since one full-disclosure path collects one sample of a given arm with constant probability, using many full-disclosure paths “in parallel” ensures that sufficiently many samples of this arm are collected.

**Theorem 5** The two-level policy with parameter \( T_1 = \tilde{O}(T^{2/3}) \) achieves regret \( \tilde{O}_K(T^{2/3}) \).

**Remark 6** For a constant \( K \), the number of arms, we match the optimal regret rate for non-adaptive multi-armed bandit algorithms. If the gap parameter \( \Delta \) is known to the principal, then (for an appropriate tuning of parameter \( T_1 \)) we can achieve regret \( \text{Reg}(T) \leq O_K(\log(T) \cdot \Delta^{-2}) \).

The proofs can be found in Appendix B. One important quantity is the expected number of samples of a given arm \( a \) collected by a full-disclosure path \( S \) of length \( L_F^{\text{FDP}} \) (i.e., present in the subhistory \( H_S \)). Indeed, this number, denoted \( N_{K,a}^{\text{FDP}} \), is the same for all such paths.

**Lemma 7** Suppose the info-graph contains \( T_1 \) full-disclosure paths of \( L_F^{\text{FDP}} \) rounds each. Let \( N_a \) be the number of samples of arm \( a \) collected by all paths. Then with probability at least \( 1 - \delta \),

\[
|N_a - N_{K,a}^{\text{FDP}}| \leq L_F^{\text{FDP}} \cdot \sqrt{T_1 \log(2K/\delta)/2}, \quad \text{for all } a \in A.
\]
4. Adaptive exploration with a three-level disclosure policy

The two-level policy from the previous section implements the explore-then-exploit paradigm using a basic design with parallel full-disclosure paths. The next challenge is to implement adaptive exploration, and go below the $T^{2/3}$ barrier. We accomplish this using a construction that adds a middle level to the info-graph. This construction also provides intuition for the main result, the multi-level construction presented in the next section. For simplicity, we assume $K = 2$ arms.

For the sake of intuition, consider the framework of bandit algorithms with limited adaptivity Perchet et al. (2016). Suppose a bandit algorithm outputs a distribution $p_t$ over arms in each round $t$, and the arm $a_t$ is then chosen independently from $p_t$. This distribution can change only in a small number of rounds, called adaptivity rounds, that need to be chosen by the algorithm in advance. A single round of adaptivity corresponds to explore-then-exploit paradigm. Our goal here is to implement one extra adaptivity round, and this is what the middle level accomplishes.

**Construction 8** The three-level policy is defined as follows (see Figure 2). The info-graph consists of three levels: the first two correspond to exploration, and the third implements exploitation. Like in the two-level policy, the first level consists of multiple full-disclosure paths of length $L_{K}^{FDP}$ each, and each agent $t$ in the exploration level sees full history from exploration. The middle level consists of $\sigma$ disjoint subsets of $T_2$ agents each, called second-level groups. Each second-level group $G$ has the following property:

\[
\text{all nodes in } G \text{ are connected to the same nodes outside of } G, \text{ but not to one another.} \tag{2}
\]

The full-disclosure paths in the first level are also split into $\sigma$ disjoint subsets, called first-level groups. Each first-level group consists of $T_1$ full-disclosure paths, for the total of $T_1 \cdot \sigma \cdot L_{K}^{FDP}$ rounds in the first layer. There is a 1-1 correspondence between first-level groups $G$ and second-level groups $G'$, whereby each agent in $G'$ observes the full history from exploration. The middle level consists of $\sigma$ disjoint subsets of $T_2$ agents each, called second-level groups. Each second-level group $G$ has the following property:

\[
\text{all nodes in } G \text{ are connected to the same nodes outside of } G, \text{ but not to one another.} \tag{2}
\]

The full-disclosure paths in the first level are also split into $\sigma$ disjoint subsets, called first-level groups. Each first-level group consists of $T_1$ full-disclosure paths, for the total of $T_1 \cdot \sigma \cdot L_{K}^{FDP}$ rounds in the first layer. There is a 1-1 correspondence between first-level groups $G$ and second-level groups $G'$, whereby each agent in $G'$ observes the full history from the corresponding group $G$. More formally, agent in $G'$ is connected to the last node of each full-disclosure path in $G$. In other words, this agent receives message $\mathcal{H}_S$, where $S$ is the set of all rounds in $G$.

The key idea is as follows. Consider the gap parameter $\Delta = |\mu_1 - \mu_2|$. If it is large, then each first-level group produces enough data to determine the best arm with high confidence, and so each agent in the upper levels chooses the best arm. If $\Delta$ is small, then due to anti-concentration each arm gets “lucky” within at least once first-level group, in the sense that it appears much better than the other arm based on the data collected in this group (and therefore this arm gets explored by the corresponding second-level group). To summarize, the middle level exploits if the gap parameter is large, and provides some more exploration if it is small.

**Theorem 9** For two arms, the three-level policy achieves regret $\text{Reg}(T) \leq O \left( T^{4/7} \log T \right)$. This is achieved with parameters $T_1 = T^{4/7} \log^{-1/7}(T)$, $\sigma = 2^{10} \log(T)$, and $T_2 = T^{6/7} \log^{-5/7}(T)$.

Let us sketch the proof of this theorem; the full proof can be found in Appendix C.

**The “good events”**. We establish four “good events” which occur with high probability.

(event1) Exploration in Level 1: Every first-level group collects $\Omega(T_1)$ samples of each arm.

(event2) Concentration in Level 1: Within each first-level group, empirical mean rewards of each arm $a$ concentrate around $\mu_a$. 

(event 3) Anti-concentration in Level 1: For each arm, some first-level subgroup collects data which makes this arm look much better than its actual mean and other arms look worse than their actual means.

(event 4) Concentration in prefix: Empirical mean reward of each arm $a$ concentrates around $\mu_a$ in any prefix of its pulls (ensuring accurate reward estimates in exploitation.)

The analysis of these events applies Chernoff Bounds to a suitable version of “reward tape” (see the definition of “reward tape” in Section 2). For example, event 2 considers a reward tape restricted to a given first-level group.

**Case analysis.** We now proceed to bound the regret conditioned on the four “good events”. W.l.o.g., assume $\mu_1 \geq \mu_2$. We break down the regret analysis into four cases, based on the magnitude the gap parameter $\Delta = \mu_1 - \mu_2$. As a shorthand, $\text{conf}(n) = \sqrt{\log(T)/n}$. In words, this is a confidence term, up to constant factors, for $n$ independent random samples. The simplest case is very small gap, which trivially yields an upper bound on regret.

**Claim 10 (Negligible gap)** If $\Delta \leq 3\sqrt{2} \cdot \text{conf}(T_2)$ then $\text{Reg}(T) \leq O(T^{4/7} \log^{6/7}(T))$.

The second case is when $\Delta$ is sufficiently large, so that the data collected in any first-level group suffices to determine the best arm. The proof follows from event 1 and event 2.

**Lemma 11 (Large gap)** If $\Delta \geq 4\sum_{a \in A} \text{conf}\left(N_{K,a}^{\text{FDP}} \cdot T_1\right)$ then all agents in the second and the third levels pull arm 1.

In the medium gap case, the data collected in a given first-level group is no longer guaranteed to determine the best arm. However, agents in the third level see the history of not only one but all first-level groups and the data collected by all first-level groups enables agents in the third level to correctly identify the best arm.
Lemma 12 (Medium gap) All agents pull arm 1 in the third level, when $\Delta$ satisfies

$$\Delta \in \left[4 \sum_{a \in A} \text{conf} \left(\sigma \cdot N_{K,a}^{\text{FDP}} \cdot T_1\right), \ 4 \sum_{a \in A} \text{conf} \left(\sigma \cdot N_{K,a}^{\text{FDP}} \cdot T_1\right)\right]$$

Finally, the small gap case, when $\Delta$ is between $\tilde{\Omega}(\sqrt{1/T_2})$ and $\tilde{O}(\sqrt{1/(\sigma T_1)})$ is more challenging since even aggregating the data from all $\sigma$ first-level groups is not sufficient for identifying the best arm. We need to ensure that both arms continue to be explored in the second level. To achieve this, we leverage event 3, which implies that each arm $a$ has a first-level group $s_a$ where it gets “lucky”, in the sense that its empirical mean reward is slightly higher than $\mu_a$, while the empirical mean reward of the other arm is slightly lower than its true mean. Since the deviations are in the order of $\Omega(\sqrt{1/T_1})$, and Assumption 1 guarantees the agents’ reward estimates are also within $\Omega(\sqrt{1/T_1})$ of the empirical means, the sub-history from this group $s_a$ ensures that all agents in the respective second-level group prefer arm $a$. Therefore, both arms are pulled at least $T_2$ times in the second level, which in turn gives the following guarantee:

Lemma 13 (Small gap) All agents pull arm 1 in the third level, when $\Delta$ satisfies

$$\Delta \in \left(3\sqrt{2} \cdot \text{conf} \left(T_2\right), \ 4 \sum_{a \in A} \text{conf} \left(\sigma \cdot N_{K,a}^{\text{FDP}} \cdot T_1\right)\right)$$

Wrapping up: proof of Theorem 9. In negligible gap case, the stated regret bound holds regardless of what the algorithm does. In the large gap case, the regret only comes from the first level, so it is upper-bounded by the total number of agents in this level, which is $\sigma \cdot L_{\text{FDP}} \cdot T_1 = O(T^{4/7} \log T)$. In both intermediate cases, it suffices to bound the regret from the first and second levels, so

$$\text{Reg}(T) \leq (\sigma T_1 \cdot L_{K,a}^{\text{FDP}} + \sigma T_2) \cdot 4 \sum_{a \in A} \text{conf} \left(N_{K,a}^{\text{FDP}} \cdot T_1\right),$$

which is at most $O(T^{4/7} \log^{6/7}(T))$. Therefore, we obtain the stated regret bound in all cases.

5. $\tilde{O}(\sqrt{T})$ regret with $L$-level policy

We give an overview of how we extend our three-level policy to a more adaptive $L$-level policy for $L > 3$ in order to achieve a regret rate of $O_K(\sqrt{T} \text{polylog}(T))$. We provide two such policies. The first policy achieves the root-$T$ regret rate with $O(\log \log T)$ levels.

Theorem 14 For any $L > 3$, there is an $L$-level disclosure policy with regret $O_K \left( T^{2L-1/(2L-1)} \cdot \text{polylog}(T) \right)$. In particular, there is a $O(\log \log(T))$-level recommendation policy with regret $O_K(T^{1/2} \cdot \text{polylog}(T))$.

Our second policy achieves an instance-dependent regret guarantee. This policy has the same info-graph structure as the first one in Theorem 14, but requires a higher number of levels $L = O(\log(T/\log \log T))$ and different group sizes. We will bound its regret as a function of the gap parameter $\Delta$ even though the construction of the policy does not depend on $\Delta$. In particular, this regret bound outperforms the one in Theorem 14 when $\Delta$ is much bigger than $T^{-1/2}$. It also has the desirable property that the policy does not withhold too much information from agents—any agent $t$ observes a good fraction of history in previous rounds.

Theorem 15 There exists an $O(\log(T)/\log \log(T))$-level policy such that for every multi-armed bandit instance with gap parameter $\Delta$, the policy has regret $O_K(\min(1/\Delta, T^{1/2}) \cdot \text{polylog}(T))$. Moreover, under this policy, each agent $t$ observes a subhistory of size at least $\Omega(t/\text{polylog}(T))$.
For constant number of arms, this result matches the optimal regret rate (1) for stochastic bandits, up to logarithmic factors. Below we present the main techniques in our solution, and the full proofs of Theorem 14 and Theorem 15 will be deferred to Appendix D. Similarly as Section 4, we first prove them in the case of 2 arms (Theorem 25 and Corollary 28). We then extend them to the case of constant number of arms (Theorem 29).

A natural idea to extend the three-level policy is to insert more levels as multiple “check points”, so the policy can incentivize the agents to perform more adaptive exploration. However, we need two key modifications in the info-graph to accommodate some new challenges. We first informally describe our techniques for the two-arm case.

**Interlacing connections between levels.** A tempted approach to generalize the three-level policy is to build an \( L \)-level info-graph with the structure of a \( \sigma \)-ary tree: for every \( l \in \{2, \ldots, L\} \), each \( l \)-level group observes the sub-history from a disjoint set \( \sigma \) groups in level \((l-1)\). The disjoint sub-histories observed by all the groups in level \( l \) are independent, and under the small gap regime (similar to Lemma 13) it ensures that each arm \( a \) has a “lucky” \( l \)-level group of agents that only pull \( a \). This “lucky” property is crucial for ensuring that both arms will be explored in level \( l \).

However, in this construction, the first level will have \( \sigma^{L-1} \) groups, which introduces a multiplicative factor of \( \sigma^{\Omega(L)} \) in the regret rate. The exponential dependence in \( L \) will heavily limit the adaptivity of the policy, and prevents having the number of levels for obtaining the result in Theorem 15. To overcome this, we will design an info-graph structure such that the number of groups at each level stays as \( \sigma^2 = \Theta(\log^2(T)) \).

We will leverage the following key observation: in order to maintain the “lucky” property, it suffices to have \( \Theta(log(T)) \) \( l \)-th level groups that observe disjoint sub-histories that take place in level \((l-1)\). Moreover, as long as the group size in levels lower than \((l-1)\) are substantially smaller than group size of level \( l-1 \), the “lucky” property does not break even if different groups in level \( l \) observe overlapping sub-history from levels \( \{1, \ldots, l-2\} \).

This motivates the following interlacing connection structure between levels. For each level in the info-graph, there are \( \sigma^2 \) groups for some \( \sigma = \Theta(\log(T)) \). The groups in the \( l \)-th level are labeled as \( G_{l,u,v} \) for \( u, v \in [\sigma] \). For any \( l \in \{2, \ldots, L\} \) and \( u, v, w \in [\sigma] \), agents in group \( G_{l,u,v} \) see the history of agents in group \( G_{l-1,v,u} \) (and by transitivity all agents in levels below \( l-1 \)). See Figure 3 for the simple case with \( \sigma = 2 \).

Two observations are in order. (i) Consider level \((l-1)\) and fix the last group index to be \( v \), and consider the set of groups \( G_{l-1,v} = \{G_{l-1,i,v} \mid i \in [\sigma]\} \) (e.g. \( G_{l-1,1,1} \) and \( G_{l-1,2,1} \) circled in red in the Figure 3). The agents in any group of \( G_{l-1,v} \) observe the same sub-history. As a result, if the empirical mean of arm \( a \) is sufficiently high in their shared sub-history, then all groups in \( G_{l-1,v} \) will become “lucky” for \( a \). (ii) Every agent in level \( l \) observes the sub-history from \( \sigma \) \((l-1)\)-th level groups, each of which belonging to a different set \( G_{l-1,v} \). Thus, for each arm \( a \), we just need one set of groups \( G_{l-1,v} \) in level \((l-1)\) to be “lucky” for \( a \) and then all agents in level \( l \) will see sufficient arm \( a \) pulls.

**Amplifying groups for boundary cases.** Recall in the three-level policy, the medium gap case (Lemma 12) corresponds to the case where the gap \( \Delta \) is between \( \Omega \left(\sqrt{1/T_1}\right) \) and \( O \left(\sqrt{\log(T)/T_1}\right) \). This is a boundary case since \( \Delta \) is neither large enough to conclude that with high probability agents in both the second level and the third level all pull the best arm, nor small enough to conclude that both arms are explored enough times in the second level (due to anti-concentration). In this case, we need to ensure that agents in the third level can eliminate the inferior arm. This issue is eas-
ily resolved in the three-level policy since the agents in the third level observe the entire first-level history, which consists of $\Omega(T_1 \log(T))$ pulls of each arm and provides sufficiently accurate reward estimates to distinguish the two arms.

In the $L$-level policy, such boundary cases occur for each intermediate level $l \in \{2, \ldots, l-1\}$, but the issue mentioned above does not get naturally resolved since the ratios between the upper and lower bounds of $\Delta$ increase from $\Theta(\sqrt{\log(T)})$ to $\Theta(\log(T))$, and it would require more observations from level $(l-2)$ to distinguish two arms at level $l$. The reason for this larger disparity is that, except the first level, our guarantee on the number of pulls of each arm is no longer tight. For example, as shown in Figure 3, when we talk about having enough arm $a$ pulls in the history observed by agents in $G_{l,1,1}$, it could be that only agents in group $G_{l-1,1,1}$ are pulling arm $a$ and it also could be that most agents in groups $G_{l-1,1,1}, G_{l-1,1,2}, \ldots, G_{l-1,1,\sigma}$ are pulling arm $a$. Therefore our estimate of the number of arm $a$ pulls can be off by an $\sigma = \Theta(\log(T))$ multiplicative factor. This ultimately makes the boundary cases harder to deal with.

We resolve this problem by introducing an additional type of amplifying groups, called $\Gamma$-groups. For each $l \in [L], u, v \in [\sigma]$, we create a $\Gamma$-group $\Gamma_{l,u,v}$. Agents in $\Gamma_{l,u,v}$ observe the same history as the one observed by agents in $G_{l,u,v}$ and the number of agents in $\Gamma_{l,u,v}$ is $\Theta(\log(T))$ times the number of agents in $G_{l,u,v}$. The main difference between $G$-groups and $\Gamma$-groups is that the history of $\Gamma$-groups in level $l$ is not sent to agents in level $l+1$ but agents in higher levels. When we are in the boundary case in which we don’t have good guarantees about the $l+1$ level agents’ pulls, the new construction makes sure that agents in levels higher than $l+1$ get to see enough pulls of each arm and all pull the best arm.

Figure 3: Interlacing connections between levels for the $L$-level policy.
References
Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, November 2013. A preliminary version has been available as a working paper since 2007.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. on Computing (SICOMP)*, 43(1):194–230, 2014. Preliminary version in *10th ACM EC*, 2009.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. of Computing (SICOMP)*, 43(1):194–230, 2014. Preliminary version in *10th ACM EC*, 2009.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. *J. of the ACM*, 62(2):10, 2015. Subsumes the conference papers in *ACM EC 2010* and *ACM EC 2013*.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. of the ACM*, 65(3), 2018. Preliminary version in *FOCS 2013*.

Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic recommendation systems. In *16th ACM Conf. on Electronic Commerce (EC)*, 2016.

Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 115–131, 2014.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *CoRR arXiv:1704.09011*, 2018. Working paper.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010. Preliminary versions have been available since 2006.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57:1407–1420, 2009.

Kostas Bimpikis, Yiangos Papanastasiou, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64:1477–1973, 2018.

Patrick Bolton and Christopher Harris. Strategic Experimentation. *Econometrica*, 67(2):349–374, 1999.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multiarmed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1), 2012.

Yeon-Koo Che and Johannes Hörner. Optimal design for social learning. *Quarterly Journal of Economics*, 2018. Forthcoming. First published draft: 2013.
Bangrui Chen, Peter I. Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In *Conf. on Learning Theory (COLT)*, pages 798–818, 2018.

Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 99–106, 2009.

Peter Frazier, David Kempe, Jon M. Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *ACM Conf. on Economics and Computation (ACM EC)*, pages 5–22, 2014.

Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: multi-armed bandits with endogenous arms. In *Innovations in Theoretical Computer Science Conf. (ITCS)*, pages 233–246, 2013.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *J. of Artificial Intelligence Research*, 55:317–359, 2016. Preliminary version appeared in *ACM EC 2014*.

Sham M. Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.

Sampath Kannan, Michael J. Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC ’17, Cambridge, MA, USA, June 26-30, 2017*, pages 369–386, 2017. doi: 10.1145/3033274.3085154. URL http://doi.acm.org/10.1145/3033274.3085154.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Godfrey Keller, Sven Rady, and Martin Cripps. Strategic Experimentation with Exponential Bandits. *Econometrica*, 73(1):39–68, 2005.

Robert D. Kleinberg and Frank T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2003.

Robert D. Kleinberg, Bo Waggoner, and E. Glen Weyl. Descending price optimally coordinates search. Working paper, 2016. Preliminary version in *ACM EC 2016*.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the wisdom of the crowd. *J. of Political Economy*, 122:988–1012, 2014. Preliminary version in *ACM EC 2014*.

Annie Liang and Xiaosheng Mu. Overabundant information and learning traps. In *ACM Conf. on Economics and Computation (ACM EC)*, pages 71–72, 2018.

Annie Liang, Xiaosheng Mu, and Vasilis Syrgkanis. Optimal and myopic information acquisition. In *ACM Conf. on Economics and Computation (ACM EC)*, pages 45–46, 2018.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *15th ACM Conf. on Economics and Computation (ACM EC)*, 2015.
I

NCENTIVIZING EXPLORATION WITH UNBIASED HISTORIES

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In 16th ACM Conf. on Economics and Computation (ACM EC), 2016.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. Working paper, 2018a. Available at https://arxiv.org/abs/1502.04147. Preliminary version in ACM EC 2015. To be published in Operations Research.

Yishay Mansour, Aleksandrs Slivkins, and Steven Wu. Competing bandits: Learning under competition. In 9th Innovations in Theoretical Computer Science Conf. (ITCS), 2018b.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. Ann. Statist., 44(2):660–681, 04 2016. doi: 10.1214/15-AOS1381. URL https://doi.org/10.1214/15-AOS1381.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In Conf. on Learning Theory (COLT), pages 1724–1738, 2018.

Sven Schmit and Carlos Riquelme. Human interaction with recommendation systems. In Intl. Conf. on Artificial Intelligence and Statistics (AISTATS), pages 862–870, 2018.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In 22nd Intl. World Wide Web Conf. (WWW), pages 1167–1178, 2013.

Appendix A. Tools from Probability: concentration and anti-concentration

We use standard tools for concentration and anti-concentration, stated below.

Theorem 16 (Chernoff Bounds) Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [0, 1]$ for all $i$. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ denote their empirical mean. Then
\[
\Pr[|\bar{X} - \mathbb{E}[\bar{X}]| > \varepsilon] \leq 2 \exp(-2n\varepsilon^2).
\]

Theorem 17 (Berry-Esseen Theorem) Let $X_1, \ldots, X_n$ be i.i.d. variables with $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] = \sigma^2 > 0$ and $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^3] = \rho < \infty$. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Let $F_n$ be the cumulative distribution function of $\frac{\bar{X} - \mathbb{E}[\bar{X}]}{\sigma \sqrt{n}}$ and $\Phi$ be the cumulative distribution function of the standard normal distribution. For all $x$ and $n$,
\[
|F_n(x) - \Phi(x)| \leq \frac{\rho}{2\sigma^3 \sqrt{n}}.
\]

Appendix B. Proofs from Section 3

Proof [Proof of Lemma 4] Fix any arm $a$. Let $L^{FDP}_K = (K - 1) \cdot N_{est} + 1$ and $p^{FDP}_K = (1/3) L^{FDP}_K$. We will condition on the event that all the realized rewards in $L^{FDP}_K$ rounds are 0, which occurs with probability at least $p^{FDP}_K$ under Assumption 1. In this case, we want to show that arm $a$ is pulled at least once. We prove this by contradiction. Suppose arm $a$ is not pulled. By the pigeonhole
principle, we know that there is some other arm \( a' \) that is pulled at least \( N_{\text{est}} + 1 \) rounds. Let \( t \) be the round in which arm \( a' \) is pulled exactly \( N_{\text{est}} + 1 \) times. By Assumption 1, we know
\[
\hat{\mu}_{a'}^{t} \leq 0 + C_{\text{est}} / \sqrt{N_{\text{est}}} \leq C_{\text{est}} < 1 / 3.
\]
On the other hand, we have \( \hat{\mu}_{a}^{t} \geq 1 / 3 > \hat{\mu}_{a'}^{t} \). This contradicts with the fact that in round \( t \), arm \( a' \) is pulled, instead of arm \( a \).

**Proof** [Proof of Theorem 5] We will set \( T_1 \) later in the proof, depending on whether the gap parameter \( \Delta \) is known. For now, we just need to know we will make \( T_1 \geq 4(L_{K}^{\text{FDP}})^2 \log(T) \). Since this policy is agnostic to the indices of the arms, we assume w.l.o.g. that arm 1 has the highest mean.

The first \( T_1 \cdot L_{K}^{\text{FDP}} \) rounds will get total regret at most \( T_1 \cdot L_{K}^{\text{FDP}} \). We focus on bounding the regret from the second level of \( T - T_1 \cdot L_{K}^{\text{FDP}} \) rounds. We consider the following two events. We will first bound the probability that both of them happen and then we will show that they together imply upper bounds on \( |\hat{\mu}_{a}^{t} - \mu_{a}| \)’s for any agent \( t \) in the second level. Recall \( \hat{\mu}_{a}^{t} \) is the estimated mean of arm \( a \) by agent \( t \) and agent \( t \) picks the arm with the highest \( \hat{\mu}_{a}^{t} \).

Define \( W_1^a \) to be the event that the number of arm \( a \) pulls in the first level is at least \( N_{K,a}^{\text{FDP}} T_1 - L_{K}^{\text{FDP}} \sqrt{T_1 \log(T)} \). As long as we set \( T_1 \geq 4(L_{K}^{\text{FDP}})^2 \log(T) \), this implies that the number of arm \( a \) pulls is then at least \( N_{K,a}^{\text{FDP}} T_1 / 2 \). Define \( W_1 \) to be the intersection of all these events (i.e. \( W_1 = \bigcap_{a} W_1^a \)). By Lemma 7, we have
\[
\Pr[W_1] \geq 1 - \frac{K}{T^2} \geq 1 - \frac{1}{T}.
\]
Next, we show that the empirical mean of each arm \( a \) is close to the true mean. To facilitate our reasoning, let us imagine there is a tape of length \( T \) for each arm \( a \), with each cell containing an independent draw of the realized reward from the distribution \( D_a \). Then for each arm \( a \) and any \( \tau \in [T] \), we can think of the sequence of the first \( \tau \) realized rewards of \( a \) coming from the prefix of \( \tau \) cells in its reward tape. Define \( W_2^{a,\tau} \) to be the event that the empirical mean of the first \( \tau \) realized rewards in the tape of arm \( a \) is at most \( \sqrt{2 \log(T) / \tau} \) away from \( \mu_a \). Define \( W_2 \) to be the intersection of these events (i.e. \( \bigcap_{a, \tau \in [T]} W_2^{a,\tau} \)). By Chernoff bound,
\[
\Pr[W_2^{a,\tau}] \geq 1 - 2 \exp(-4 \log(T)) \geq 1 - 2/T^4.
\]
By union bound,
\[
\Pr[W_2] \geq 1 - KT \cdot \frac{2}{T^4} \geq 1 - \frac{2}{T}.
\]
By union bound, we know \( \Pr[W_1 \cap W_2] \geq 1 - 3/T \). For the remainder of the analysis, we will condition on the event \( W_1 \cap W_2 \).

For any arm \( a \) and agent \( t \) in the second level, by \( W_1 \) and \( W_2 \), we have
\[
|\hat{\mu}_{a}^{t} - \mu_{a}| \leq \sqrt{\frac{2 \log(T)}{N_{K,a}^{\text{FDP}} T_1 / 2}}.
\]
By \( W_1 \) and Assumption 1, we have
\[
|\hat{\mu}_{a}^{t} - \hat{\mu}_{a}^{t}| \leq \frac{C_{\text{est}}}{\sqrt{N_{K,a}^{\text{FDP}} T_1 / 2}}.
\]
Therefore,
\[ |\mu_a^t - \mu_a| \leq \sqrt{\frac{2 \log(T)}{N_{FDP}^K T_1/2}} + \frac{C_{est}}{\sqrt{N_{FDP}^K T_1/2}} \leq 3 \frac{\log(T)}{p_{FDP}^K T_1}. \]

So the second-level agents will pick an arm \( a \) which has \( \mu_a \) at most \( 6 \sqrt{\frac{\log(T)}{p_{FDP}^K T_1}} \) away from \( \mu_1 \). To sum up, the total regret is at most
\[ T_1 \cdot L_{FDP}^K + T \cdot (1 - \Pr[W_1 \cap W_2]) + T \cdot 6 \sqrt{\frac{\log(T)}{p_{FDP}^K T_1}}. \]

By setting \( T_1 = T^{2/3} \log(T)^{1/3} \), we get regret \( O(T^{2/3} \log(T)^{1/3}) \).

\[ \square \]

**Appendix C. Missing proofs from Section 4**

**C.1. Events**

The following lemmas can be derived from combining Lemma 7 and union bound.

**Lemma 18 (Concentration of first-level number of pulls.)** Let \( W_1 \) be the event that for all groups \( s \in [\sigma] \) and arms \( a \in \{1, 2\} \), the number of arm \( a \) pulls in the \( s \)-th first-level group is in the range of
\[ \left[ N_{FDP}^K T_1 - L_{FDP}^K \sqrt{T_1 \log(T)}, N_{FDP}^K T_1 + L_{FDP}^K \sqrt{T_1 \log(T)} \right], \]
where \( N_{FDP}^K \) is the expected number of arm \( a \) pulls in a full-disclosure path run of length \( L_{FDP}^K \). Then \( \Pr[W_1] \geq 1 - \frac{4\sigma}{T^2} \).

**Proof** [Proof of Lemma 18] For the \( s \)-th first-level group, define \( W_1^{a,s} \) to be the event that the number of arm \( a \) pulls in the \( s \)-th first-level group is between \( N_{FDP}^K T_1 - L_{FDP}^K \sqrt{T_1 \log(T)} \) and \( N_{FDP}^K T_1 + L_{FDP}^K \sqrt{T_1 \log(T)} \). By Lemma 7
\[ \Pr[W_1^{a,s}] \geq 1 - 2 \exp(-2 \log(T)) \geq 1 - 2/T^2. \]

Let \( W_1 \) be the intersection of all these events (i.e. \( W_1 = \bigcap_{a,s} W_1^{a,s} \)). By union bound, we have
\[ \Pr[W_1] \geq 1 - \frac{4\sigma}{T^2}. \]

\[ \square \]

To state the events, it will be useful to think of a hypothetical reward tape \( T_{s,a}^1 \) of length \( T \) for each group \( s \) and arm \( a \), with each cell independently sampled from \( D_a \). The tape encodes rewards as follows: the \( j \)-th time arm \( a \) is chosen by the group \( s \) in the first level, its reward is taken from the \( j \)-th cell in this arm’s tape. The following result characterizes the concentration of the mean rewards among all consecutive pulls among all such tapes, which follows from Chernoff bound and union bound.
Lemma 19 (Concentration of empirical means in the first level) For any $\tau_1, \tau_2 \in [T]$ such that $\tau_1 < \tau_2, s \in [\sigma], \text{and } a \in \{1, 2\}$, let $W_2^{s,a,\tau_1,\tau_2}$ be the event that the mean among the cells indexed by $\tau_1, (\tau_1 + 1), \ldots, \tau_2$ in the tape $\mathcal{T}_{a,s}^1$ is at most $\sqrt{2 \log(T)} / (\tau_2 - \tau_1 + 1)$ away from $\mu_a$. Let $W_2$ be the intersection of all these events (i.e. $W_2 = \bigcap_{a,s,\tau_1,\tau_2} W_2^{s,a,\tau_1,\tau_2}$). Then
\[
\Pr[W_2] \geq 1 - \frac{4\sigma}{T^2}.
\]

Proof [Proof of Lemma 19] By Chernoff bound,
\[
\Pr[W_2^{s,a,\tau_1,\tau_2}] \geq 1 - 2\exp(-4 \log(T)) \geq 1 - \frac{4}{T^4}.
\]
By union bound, we have
\[
\Pr[W_2] \geq 1 - \frac{4\sigma}{T^2}.
\]

Our policy also relies on the anti-concentration of the empirical means in the first round. We show that for each arm $a \in \{1, 2\}$, there exists a group $s_a$ such that the empirical mean of $a$ is slightly above $\mu_a$, while the other arm $(3 - a)$ has empirical mean slightly below $\mu_{(3-a)}$. This event is crucial for inducing agents in the second level to explore both arms when the their mean rewards are indistinguishable after the first level.

Lemma 20 (Co-occurrence of high and low deviations in this first level) For any group $s \in [\sigma], \text{any arm } a, \text{let } \bar{\mu}_{a,s} \text{ be the empirical mean of the first } N_{FDP}^{\tau_1} \text{ cells in tape } \mathcal{T}_{a,s}^1. \text{ Let } W_3^{s,a,\text{high}} \text{ be the event } \bar{\mu}_{a,s} \geq \mu_a + 1/\sqrt{N_{FDP}^{\tau_1}} \text{ and let } W_3^{s,a,\text{low}} \text{ be the event that } \bar{\mu}_{a,s} \leq \mu_a - 1/\sqrt{N_{FDP}^{\tau_1}}. \text{ Let } W_3 \text{ be the event that for every } a \in \{1, 2\}, \text{there exists a group } s_a \in [\sigma] \text{ in the first level such that both } W_3^{s_a,\text{high}} \text{ and } W_3^{s_a,3-a,\text{low}} \text{ occur. Then }
\[
\Pr[W_3] \geq 1 - 2/T.
\]

Proof [Proof of Lemma 20] By Berry-Esseen Theorem and $\mu_a \in [1/3, 2/3]$, we have for any $a$,
\[
\Pr[W_3^{s,a,\text{high}}] \geq (1 - \Phi(1/2)) - \frac{5}{\sqrt{N_{FDP}^{\tau_1}}} > 1/4.
\]
The last inequality follows when $T$ is larger than some constant. Similarly we also have
\[
\Pr[W_3^{s,a,\text{low}}] > 1/4.
\]
Since $W_3^{s,a,\text{high}}$ is independent with $W_3^{s_3,3-a,\text{low}}$, we have
\[
\Pr[W_3^{s,a,\text{high}} \cap W_3^{s_3,3-a,\text{low}}] = \Pr[W_3^{s,a,\text{high}}] \cdot \Pr[W_3^{s_3,3-a,\text{low}}] > (1/4)^2 = 1/16.
\]
Notice that $(W_3^{s,a,\text{high}} \cap W_3^{s_3,3-a,\text{low}})$ are independent across different $s$’s. By union bound, we have
\[
\Pr[W_3] \geq 1 - 2(1 - 1/16)^{\sigma} \geq 1 - 2/T.
\]
Lastly, we will condition on the event that the empirical means of both arms are concentrated around their true means in any prefix of their pulls. This guarantees that the policy obtains an accurate estimate of rewards for both arms after aggregating all the data in the first two levels.

**Lemma 21 (Concentration of empirical means in the first two levels)** With probability at least $1 - \frac{4}{T^3}$, the following event $W_4$ holds: for all $a \in \{1, 2\}$ and $\tau \in [N_{T,a}]$, the empirical means of the first $\tau$ arm $a$ pulls is at most $\sqrt{\frac{2\log(T)}{\tau}}$ away from $\mu_a$, where $N_{T,a}$ is the total number of arm $a$ pulls by the end of $T$ rounds.

**Proof** [Proof of Lemma 21] For any arm $a$, let’s imagine a hypothetical tape of length $T$, with each cell independently sampled from $D_a$. The tape encodes rewards of the first two levels as follows: the $j$-th time arm $a$ is chosen in the first two levels, its reward is taken from the $j$-th cell in the tape.

Define $W_4^{a,\tau}$ to be the event that the mean of the first $t$ pulls in the tape is at most $\sqrt{\frac{2\log(T)}{\tau}}$ away from $\mu_a$. By Chernoff bound,

$$\Pr[W_4^{a,\tau}] \geq 1 - 2\exp(-4\log(T)) \geq 1 - 2/T^4.$$  

By union bound, the intersection of all these events has probability at least:

$$\Pr[W_4] \geq 1 - \frac{4}{T^3}.$$  

Let $W = \cap_{i=1}^4 W_i$ be the intersection of all 4 events. By union bound, $W$ occurs with probability $1 - O(1/T)$. Note that the regret conditioned on $W$ not occurring is at most $O(1/T) \cdot T = O(1)$, so it suffices to bound the regret conditioned on $W$.

**C.2. Case Analysis**

Now we assume the intersection $W$ of events $W_1, \cdots, W_4$ happens. We will first provide some helper lemmas for our case analysis.

**Lemma 22** For the $s$-th first-level group and arm $a$, define $\hat{\mu}_a^{1,s}$ to be the empirical mean of arm $a$ pulls in this group. If $W$ holds, then

$$|\hat{\mu}_a^{1,s} - \mu_a| \leq \sqrt{\frac{4\log(T)}{N_{K,a}^F T_1}}.$$  

**Proof** The events $W_1$ and $W_2^{a,s,1,\tau}$ for $\tau = N_{K,a}^F T_1 - L_K^F \sqrt{T_1 \log(T)}$ together imply that

$$|\hat{\mu}_a^{1,s} - \mu_a| \leq \sqrt{\frac{2\log(T)}{N_{K,a}^F T_1 - L_K^F \sqrt{T_1 \log(T)}}} \leq \sqrt{\frac{4\log(T)}{N_{K,a}^F T_1}}.$$  

The last inequality holds when $T$ is larger than some constant.
Lemma 23  For each arm $a$, define $\bar{\mu}_a$ to be the empirical mean of arm $a$ pulls in the first two levels. If $W$ holds, then

$$|\bar{\mu}_a - \mu_a| \leq \sqrt{\frac{4 \log(T)}{\sigma N_{K,a} T_1}}.$$ 

Furthermore, if there are at least $T_2$ pulls of arm $a$ in the first two levels,

$$|\bar{\mu}_a - \mu_a| \leq \sqrt{\frac{2 \log(T)}{T_2}}.$$ 

Proof  The events $W_1$ and $W_{a,\tau}$ for $\tau \geq (N_{FDP,K,a} T_1 - L_{FDP,K} \sqrt{T_1 \log(T)})/ \sigma$ together imply that

$$|\bar{\mu}_a - \mu_a| \leq \sqrt{\frac{2 \log(T)}{\sigma N_{FDP,K,a} T_1}} \leq \sqrt{\frac{4 \log(T)}{\sigma N_{FDP,K,a} T_1}}.$$ 

The last inequality holds when $T$ is larger than some constant.

Lemma 24  For the $s$-th first-level group and arm $a$, define $\bar{\mu}_{1,s,a}$ to be the empirical mean of arm $a$ pulls in this group. For each $a \in \{1, 2\}$, there exists a group $s_a$ such that

$$\bar{\mu}_{1,s,a} > \mu_a + \frac{1}{4 N_{FDP,K,a} T_1} \quad \text{and} \quad \bar{\mu}_{3-a,s,a} < \mu_{3-a} - \frac{1}{4 N_{FDP,K,3-a} T_1}.$$ 

Proof  For each $a \in \{1, 2\}$, $W_3$ implies that there exists $s_a$ such that both $W_{s_a,a,\tau,high}$ and $W_{s_a,3-a,\tau,low}$ happen. The events $W_{s_a,a,\tau,high}$, $W_1$, $W_{2,s_a,a,\tau,low}$, $N_{FDP,K,a} T_1$ for $\tau = N_{FDP,K,a} T_1 - L_{FDP,K} \sqrt{T_1 \log(T)} + 1, \ldots, N_{FDP,K,a} T_1 - 1$ and $W_{2,s_a,a,\tau,low}$, $N_{FDP,K,a} T_1$ for $\tau = N_{FDP,K,a} T_1, \ldots, N_{FDP,K,a} T_1 + L_{FDP,K} \sqrt{T_1 \log(T)}$ together imply that

$$\bar{\mu}_{a,s} > \mu_a + \frac{1}{4 \sqrt{N_{FDP,K,a} T_1}} - L_{FDP,K} \sqrt{T_1 \log(T)} + \frac{1}{4 \sqrt{N_{FDP,K,a} T_1}} \cdot \sqrt{\frac{2 \log(T)}{L_{FDP,K} \sqrt{T_1 \log(T)}}}$$

$$> \mu_a + \frac{1}{4 \sqrt{N_{FDP,K,a} T_1}}.$$ 

The second to the last inequality holds when $T$ is larger than some constant. Similarly, we also have

$$\bar{\mu}_{3-a,s} < \mu_{3-a} - \frac{1}{4 \sqrt{N_{FDP,K,3-a} T_1}}.$$ 

This completes the proof.

Now we proceed to the case analysis.
Incentivizing Exploration with Unbiased Histories

**Proof** [Proof of Lemma 11 (Large gap case)] Observe that for any group $s$ in the first level, the empirical means satisfy

$$\bar{\mu}_1^s - \bar{\mu}_2^s \geq \mu_1 - \mu_2 - \sqrt{\frac{4 \log(T)}{N_{FDP}^{K,1} T_1}} \geq \sqrt{\frac{4 \log(T)}{N_{FDP}^{K,2} T_1}}.$$

For any agent $t$ in the $s$-th second-level group, by Assumption 1, we have

$$\hat{\mu}_t^1 - \hat{\mu}_t^2 > \bar{\mu}_1^s - \bar{\mu}_2^s - \frac{C_{est}}{\sqrt{N_{FDP}^{K,1} T_1/2}} - \frac{C_{est}}{\sqrt{N_{FDP}^{K,2} T_1/2}} \geq \sqrt{\frac{4 \log(T)}{N_{FDP}^{K,1} T_1}} + \sqrt{\frac{4 \log(T)}{N_{FDP}^{K,2} T_1}}.$$

Therefore, we know agents in the $s$-th second-level group will all pull arm 1.

Now consider the agents in the third level group. Recall $\bar{\mu}_t$ is the empirical mean of arm $a$ in the history they see. We have

$$\bar{\mu}_t - \bar{\mu}_2 \geq \mu_1 - \mu_2 - \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,1} T_1}} \geq \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,2} T_1}}.$$

Similarly as above, by Assumption 1, we know $\bar{\mu}_t^1 - \bar{\mu}_2^t > 0$ for any agent $t$ in the third level. Therefore, the agents in the third-level group will all pull arm 1.

**Proof** [Proof of Lemma 12 (Medium gap case)] Recall $\bar{\mu}_a$ is the empirical mean of arm $a$ in the first two levels. We have

$$\bar{\mu}_1 - \bar{\mu}_2 \geq \mu_1 - \mu_2 - \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,1} T_1}} \geq \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,2} T_1}}.$$

For any agent $t$ in the third level, by Assumption 1, we have

$$\hat{\mu}_t^1 - \hat{\mu}_t^2 > \bar{\mu}_1 - \bar{\mu}_2 - \frac{C_{est}}{\sqrt{\sigma N_{FDP}^{K,1} T_1/2}} - \frac{C_{est}}{\sqrt{\sigma N_{FDP}^{K,2} T_1/2}} \geq \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,1} T_1}} + \sqrt{\frac{4 \log(T)}{\sigma N_{FDP}^{K,2} T_1}}.$$

So we know agents in the third-level group will all pull arm 1.

**Proof** [Proof of Lemma 13 (Small gap case)] In this case, we need both arms to be pulled at least $T_2$ rounds in the second level. For every arm $a$, consider the $s_a$-th second-level group, with $s_a$ given
by Lemma 24. We have
\[
\hat{\mu}_a^{s_a} - \mu_{3-a} - \mu_a + \frac{1}{4\sqrt{N_{K,a}^\text{FDP} T_1}} - \frac{1}{4\sqrt{N_{K,3-a}^\text{FDP} T_1}} > 0.
\]

For any agent \( t \) in the \( s_a \)-th second-level group, by Assumption 1, we have
\[
\hat{\mu}_a^t - \mu_{3-a} - \mu_a + \frac{C_{\text{est}}}{\sqrt{N_{K,a}^\text{FDP} T_1/2}} - \frac{C_{\text{est}}}{\sqrt{N_{K,3-a}^\text{FDP} T_1/2}} > 0.
\]

So we know agents in the \( s_a \)-th second-level group will all pull arm \( a \). Therefore in the first two levels, both arms are pulled at least \( T_2 \) times. Now consider the third-level. We have
\[
\bar{\mu}_1 - \bar{\mu}_2 \geq \mu_1 - \mu_2 - 2 \sqrt{\frac{2 \log(T)}{T_2}} \geq \sqrt{\frac{2 \log(T)}{T_2}}.
\]

Similarly as above, by Assumption 1, we know \( \hat{\mu}_1^t - \hat{\mu}_2^t > 0 \) for any agent \( t \) in the third level. So we know agents in the third-level group will all pull arm 1.

**Appendix D. Proofs from Section 5**

In this section, we design our \( L \)-level recommendation policy for \( L > 3 \). Similarly as Section 4, we first prove them in the case of 2 arms (Theorem 25 and Corollary 28). We then extend them to the case of constant number of arms (Theorem 29).

Now we start with the case of 2 arms. Our recommendation policy has \( L \) levels and two types of groups: \( G \)-groups and \( \Gamma \)-groups. Each level has \( \sigma^2 \) \( G \)-groups for \( \sigma = 2^{10} \log(T) \). Label the \( G \)-groups in the \( l \)-th level as \( G_{l,u,v} \) for \( u,v \in [\sigma] \). Level 2 to level \( L \) also have \( \sigma^2 \) \( \Gamma \)-groups. Label the \( \Gamma \)-groups in the \( l \)-th level as \( \Gamma_{l,u,v} \) for \( u,v \in [\sigma] \). Each first-level group \((G_{1,u,v} \) for \( u,v \in [\sigma]) \) has \( T_1 \) full-disclosure path of \( L_{K}^{\text{FDP}} \) rounds in parallel. For \( l \geq 2 \), there are \( T_l \) agents in group \( G_{l,u,v} \) and there are \( T_l (\sigma - 1) \) agents in group \( \Gamma_{l,u,v} \). We will pick \( T_1, ..., T_L \) in the proof of Theorem 25.

Finally we define the info-graph. Agents in the first level only observe the history defined in the full-disclosure path run. For agents in group \( G_{l,u,v} \) with \( l \geq 2 \), they observe all the history in the first \( l - 2 \) levels (both \( G \)-groups and \( \Gamma \)-groups) and history in group \( G_{l-1,u,v} \) for all \( v \in [\sigma] \). Agents in group \( \Gamma_{l,u,v} \) observe the same history as agents in group \( G_{l,u,v} \).
Theorem 25 The $L$-level recommendation policy gets regret $O \left( T^{2^{L-1}/(2^{L-1})} \log^2(T) \right)$ for $L \leq \log(\ln(T)/\log(\sigma^4))$. In particular, if we pick $L = \log(\ln(T)/\log(\sigma^4))$, the regret is $O(T^{1/2}\text{polylog}(T))$.

Proof Wlog we assume $\mu_1 \geq \mu_2$ as the recommendation policy is symmetric to both arms. We will set $T_t's$ later in the proof. Before that, we are only going to assume $T_t/T_{t-1} \geq \sigma^4$ for $t = 2, \ldots, L-1$ and $T_1 \geq \sigma^4$.

Similarly as the proof of Theorem 9, we start with some clean events.

- Concentration of the number of arm $a$ pulls in the first level:

For $a \in \{1, 2\}$, define $N^FDP_{K,a}$ to be the expected number of arm $a$ pulls in one run of full-disclosure path used in the first level. By Lemma 4, we know $P^FDP_K \leq N^FDP_{K,a} \leq L^FDP_K$ For group $G_{1,u,v}$, define $W_{1,a,u,v}^a$ to be the event that the number of arm $a$ pulls in this group is between $N^FDP_{K,a}T_1 - L^FDP_K \sqrt{T_1 \log(T)}$ and $N^FDP_{K,a}T_1 + L^FDP_K \sqrt{T_1 \log(T)}$. By Chernoff bound,

$$\Pr[W_{1,a,u,v}^a] \geq 1 - 2 \exp(-2 \log(T)) \geq 1 - 2/T^2.$$  

Define $W_1$ to be the intersection of all these events (i.e. $W_1 = \bigcap_{a,u,v} W_{1,a,u,v}^a$). By union bound, we have

$$\Pr[W_1] \geq 1 - 4\sigma^2 T^2.$$  

- Concentration of the empirical mean of arm $a$ pulls in the history observed by agent $t$:

For each agent $t$ and arm $a$, imagine there is a tape of enough arm $a$ pulls sampled before the recommendation policy starts and these samples are revealed one by one whenever agents in agent $t$’s observed history pull arm $a$. Define $W_{2,t,a,\tau_1,\tau_2}$ to be the event that the mean of $\tau_1$-th to $\tau_2$-th pulls in the tape is at most $\sqrt{\frac{3\log(T)}{\tau_2-\tau_1+1}}$ away from $\mu_a$. By Chernoff bound,

$$\Pr[W_{2,t,a,\tau_1,\tau_2}] \geq 1 - 2 \exp(-6 \log(T)) \geq 1 - 2/T^6.$$  

Define $W_2$ to be the intersection of all these events (i.e. $W_2 = \bigcap_{t,a,\tau_1,\tau_2} W_{2,t,a,\tau_1,\tau_2}$). By union bound, we have

$$\Pr[W_2] \geq 1 - \frac{4}{T^3}.$$  

- Anti-concentration of the empirical mean of arm $a$ pulls in the $l$-th level for $l \geq 2$:

For $2 \leq l \leq L - 1$, $u \in [\sigma]$ and each arm $a$, define $n_{l,u,a}$ to be the number of arm $a$ pulls in groups $G_{l,u,1}, \ldots, G_{l,u,\sigma}$. Define $W_{3,l,u,a,\text{high}}$ as the event that $n_{l,u,a} \geq T_l$ implies the empirical mean of arm $a$ pulls in group $G_{l,u,1}, \ldots, G_{l,u,\sigma}$ is at least $\mu_a + 1/\sqrt{n_{l,u,a}}$. Define $W_{3,l,u,a,\text{low}}$ as the event that $n_{l,u,a} \geq T_l$ implies the empirical mean of arm $a$ pulls in group $G_{l,u,1}, \ldots, G_{l,u,\sigma}$ is at most $\mu_a - 1/\sqrt{n_{l,u,a}}$.

Define $H_l$ to be random variable the history of all agents in the first $l - 1$ levels and which agents are chosen in the $l$-th level. Let $h_l$ be some realization of $H_l$. Notice that once we fix $H_l$, $n_{l,u,a}$ is also fixed.
Finally we define \( u \) independent across different \( W \). Now define \( W \) when fixing \( H \), we have
\[
\Pr[W_3^{l,u,a,high}|H_t = h_t] \geq 1 - \Phi(1/2) - \frac{5}{\sqrt{T_l}} > 1/4.
\]
Similar we also have
\[
\Pr[W_3^{l,u,a,low}|H_t = h_t] > 1/4.
\]
Since \( W_3^{l,u,a,high} \) is independent with \( W_3^{l,u,3-a,low} \) when fixing \( H \), we have
\[
\Pr[W_3^{l,u,a,high} \cap W_3^{l,u,3-a,low}|H_t = h_t] > (1/4)^2 = 1/16.
\]
Now define \( W_3^{l,a} = \bigcup_a (W_3^{l,u,a,high} \cap W_3^{l,u,3-a,low}) \). Since \((W_3^{l,u,a,high} \cap W_3^{l,u,3-a,low})\) are independent across different \( u \)'s when fixing \( H \), we have
\[
\Pr[W_3^{l,a}|H_t = h_t] \geq 1 - (1 - 1/16)^a \geq 1 - 1/T^2.
\]
Since this holds for all \( h_t \)'s, we have \( \Pr[W_3^{l,a}] \geq 1 - 1/T^2 \). Finally define \( W_3 = \bigcap_{a} W_3^{l,a} \). By union bound, we have
\[
W_3 \geq 1 - 2L/T^2.
\]
- **Anti-concentration of the empirical mean of arm \( a \) pulls in the first level:**

For first-level groups \( G_{1,u,1}, \ldots, G_{1,u,a} \) and arm \( a \), imagine there is a tape of enough arm \( a \) pulls sampled before the recommendation policy starts and these samples are revealed one by one whenever agents in these groups pull arm \( a \). Define \( W_4^{u,a,high} \) to be the event that first \( N_{\text{FDP}}^{K,a} T_1 \sigma \) pulls of arm \( a \) in the tape has empirical mean at least \( \mu_a + 1/\sqrt{N_{\text{FDP}}^{K,a} T_1 \sigma} \) and define \( W_4^{u,a,low} \) to be the event that first \( N_{\text{FDP}}^{K,a} T_1 \sigma \) pulls of arm \( a \) in the tape has empirical mean at most \( \mu_a - 1/\sqrt{N_{\text{FDP}}^{K,a} T_1 \sigma} \). By Berry-Esseen Theorem and \( \mu_a \in [1/3, 2/3] \), we have
\[
\Pr[W_4^{u,a,high}] \geq (1 - \Phi(1/2)) - \frac{5}{N_{\text{FDP}}^{K,a} T_1 \sigma} > 1/4.
\]
The last inequality follows when \( T \) is larger than some constant. Similarly we also have
\[
\Pr[W_4^{u,a,low}] > 1/4.
\]
Since \( W_4^{u,a,high} \) is independent with \( W_4^{u,3-a,low} \), we have
\[
\Pr[W_4^{u,a,high} \cap W_4^{u,3-a,low}] = \Pr[W_4^{u,a,high}] \cdot \Pr[W_4^{u,3-a,low}] > (1/4)^2 = 1/16.
\]
Now define \( W_4^{a} \) as \( \bigcup_a (W_4^{u,a,high} \cap W_4^{u,3-a,low}) \). Notice that \((W_4^{u,a,high} \cap W_4^{u,3-a,low})\) are independent across different \( u \)'s. So we have
\[
\Pr[W_4^{a}] \geq 1 - (1 - 1/16)^a \geq 1 - 1/T^2.
\]
Finally we define \( W_4 = \bigcap_{a} W_4^{a} \). By union bound,
\[
\Pr[W_4] \geq 1 - 2/T^2.
\]
By union bound, the intersection of these clean events (i.e. $\bigcap_{i=1}^{4} W_i$) happens with probability $1 - O(1/T)$. When this intersection does not happen, since the probability is $O(1/T)$, it contributes $O(1/T) \cdot T = O(1)$ to the regret.

Now we assume the intersection of clean events happens and prove upper bound on the regret.

By event $W_1$, we know that in each first-level group, there are at least $N_{\text{FDP}}^{T_1} - L_{K}^{\text{FDP}} \sqrt{T_1 \log(T)}$ pulls of arm $a$. We prove in the next claim that there are enough pulls of both arms in higher levels if $\mu_1 - \mu_2$ is small enough. For notation convenience, we set $\varepsilon_0 = 1, \varepsilon_1 = \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1}T_1\sigma}} + \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1}T_1\sigma}}$ and $\varepsilon_l = 1/(4\sqrt{T_l\sigma})$ for $l \geq 2$.

**Claim 26** For any arm $a$ and $2 \leq l \leq L$, if $\mu_1 - \mu_2 \leq \varepsilon_{l-1}$, then for any $u \in [\sigma]$, there are at least $T_l$ pulls of arm $a$ in groups $G_{l,u,1}, G_{l,u,2}, ..., G_{l,u,\sigma}$ and there are at least $T_l\sigma(\sigma - 1)$ pulls of arm $a$ in the $l$-th level $\Gamma$-groups.

**Proof** We are going to show that for each $l$ and arm $a$ there exists $u_a$ such that agents in groups $G_{l,1,u_a}, ..., G_{l,\sigma,u_a}$ and $\Gamma_{l,1,u_a}, ..., \Gamma_{l,\sigma,u_a}$ pull arm $a$. This suffices to prove the claim.

We prove the above via induction on $l$. We start by the base case when $l = 2$. For each arm $a$, $W_4$ implies there exists $u_a$ such that $W_4^{u_a,a,\text{high}}$ and $W_4^{u_a,3-a,\text{low}}$ happen. For an agent $t$ in groups $G_{2,1,u_a}, ..., G_{2,\sigma,u_a}$ and $\Gamma_{2,1,u_a}, ..., \Gamma_{2,\sigma,u_a}$, $W_4^{u_a,a,\text{high}}, W_4^{u_a,3-a,\text{low}}$ and $W_2$ together imply that

$$\hat{\mu}_a^t \geq \mu_a + \frac{1}{N_{\text{FDP}}^{T_1\sigma}} \cdot \frac{1}{\sqrt{N_{\text{FDP}}^{T_1\sigma}} - L_{K}^{\text{FDP}} \sqrt{T_1 \log(T)}} \cdot \left( N_{\text{FDP}}^{T_1\sigma} + L_{K}^{\text{FDP}} \sqrt{T_1 \log(T)} \right) \cdot \frac{3 \log(T)}{L_{K}^{\text{FDP}} \sqrt{T_1 \log(T)}}$$

$$\geq \mu_a + \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma}.$$

The second last inequality holds when $T$ is larger than some constant. Similarly, we also have

$$\hat{\mu}_{3-a}^t \leq \mu_{3-a} - \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma}.$$

Then we have

$$\hat{\mu}_a^t - \hat{\mu}_{3-a}^t \geq \mu_a - \mu_{3-a} + \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma} - \frac{1}{4\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma}$$

$$\geq -\varepsilon_1 + \frac{1}{8\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma} + \frac{1}{8\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma}$$

$$\geq \frac{1}{8\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma} + \frac{1}{8\sqrt{N_{\text{FDP}}^{T_1\sigma}} T_1\sigma}.$$
By Assumption 1, we have

\[
\mu_t^a - \hat{\mu}_3^a > \mu_t^a - \hat{\mu}_3^a - \frac{C_{est}}{\sqrt{N_{K,a} T_1 \sigma^2/2}} - \frac{C_{est}}{\sqrt{N_{K,3-a} T_1 \sigma^2/2}} > \frac{1}{8 \sqrt{N_{K,a} T_1 \sigma}} + \frac{1}{8 \sqrt{N_{K,3-a} T_1 \sigma}} - \frac{C_{est}}{\sqrt{N_{K,a} T_1 \sigma^2/2}} - \frac{C_{est}}{\sqrt{N_{K,3-a} T_1 \sigma^2/2}} > 0.
\]

The last inequality holds since \(C_{est}\) is a small enough constant defined in Assumption 1. Therefore we know agents in groups \(G_{2,1,u_a}, \ldots, G_{2,\sigma,u_a}\) and \(\Gamma_{2,1,u_a}, \ldots, \Gamma_{2,\sigma,u_a}\) all pull arm \(a\).

Now we consider the case when \(l > 2\) and assume the claim is true for smaller \(l\)’s. For each arm \(a\), \(W_3\) implies that there exists \(u_a\) such that \(W_3^{l-1,u_a,a,high}\) and \(W_3^{l-1,u_a,a,low}\) happen. Recall \(n^{l-1,u_a,a}\) is the number of arm \(a\) pulls in groups \(G_{l-1,u_a,1}, \ldots, G_{l-1,u_a,\sigma}\). The induction hypothesis implies that \(n^{l-1,u_a,a} \geq T_{l-1}\). \(W_3^{l-1,u_a,a,high}\) together with \(n^{l-1,u_a,a} \geq T_{l-1}\) implies that the empirical mean of arm \(a\) pulls in group \(G_{l-1,u_a,1}, \ldots, G_{l-1,u_a,\sigma}\) is at least \(\mu_a + 1/\sqrt{n^{l-1,u_a,a}}\). For any agent \(t\) in groups \(G_{l-1,u_a,1}, \ldots, G_{l-1,u_a,\sigma}\) and all groups in levels below level \(l - 1\), notice that the groups in the first \(l - 2\) levels have at most \((T_1 L_{K,FDP} + T_2 + \cdots + T_{l-2})\) agents. By \(W_2\), we have

\[
\mu_t^a \geq \mu_a + \left(\frac{1}{n^{l-1,u_a,a}} \cdot \frac{1}{\sqrt{n^{l-1,u_a,a}}} - (T_1 L_{K,FDP} + T_2 + \cdots + T_{l-2})\sigma^2 \cdot \frac{1}{(T_1 L_{K,FDP} + T_2 + \cdots + T_{l-2})\sigma^2}\right) > \mu_a + \frac{1}{4 \sqrt{n^{l-1,u_a,a}}}
\]

The third last inequality holds when \(T\) larger than some constant. Similarly, we also have

\[
\hat{\mu}_3^a < \mu_3^a - \frac{1}{4 \sqrt{n^{l-1,u_a,3-a}}}
\]

Then we have

\[
\hat{\mu}_3^a - \mu_3^a > \mu_a - \mu_3^a + \frac{1}{4 \sqrt{n^{l-1,u_a,a}}} + \frac{1}{4 \sqrt{n^{l-1,u_a,3-a}}} \geq -\epsilon_{l-1} + \frac{1}{8 \sqrt{n^{l-1,u_a,a}}} + \frac{1}{8 \sqrt{n^{l-1,u_a,3-a}}} \geq \frac{1}{8 \sqrt{n^{l-1,u_a,a}}} + \frac{1}{8 \sqrt{n^{l-1,u_a,3-a}}}
\]
The last inequality holds because \( n^l-1,u_a,a \) and \( n^l-1,u_a,3-a \) are at most \( T_{l-1} \). By Assumption 1, we have

\[
\hat{\mu}_a^l - \hat{\mu}_{a-3}^l > \hat{\mu}_a^l - \hat{\mu}_{a-3}^l - \frac{C_{\text{est}}}{\sqrt{n^l-1,u_a,a}} - \frac{C_{\text{est}}}{\sqrt{n^l-1,u_a,3-a}}
\]

\[
> \frac{1}{8\sqrt{n^l-1,u_a,a}} + \frac{1}{8\sqrt{n^l-1,u_a,3-a}} - \frac{C_{\text{est}}}{\sqrt{n^l-1,u_a,a}} - \frac{C_{\text{est}}}{\sqrt{n^l-1,u_a,3-a}}
\]

\[
> 0.
\]

The last inequality holds since \( C_{\text{est}} \) is a small enough constant defined in Assumption 1. Therefore agents in groups \( G_{l,1,u_a}, \ldots, G_{l,\sigma,u_a} \) and \( \Gamma_{l,1,u_a}, \ldots, \Gamma_{l,\sigma,u_a} \) all pull arm \( a \).

**Claim 27** For any \( 2 \leq l \leq L \), if \( \varepsilon_{l-1} \sigma \leq \mu_1 - \mu_2 < \varepsilon_{l-2} \sigma \), there are no pulls of arm 2 in groups with level \( l, \ldots, L \).

**Proof** We argue in 2 cases \( \varepsilon_{l-1} \sqrt{\sigma} \leq \mu_1 - \mu_2 < \varepsilon_{l-2} \) for \( l > 2 \) and \( \varepsilon_{l-2} \leq \mu_1 - \mu_2 \leq \varepsilon_{l-2} \sqrt{\sigma} \) for \( l > 2 \). Since our recommendation policy’s first level is slightly different from other levels, we need to argue case \( \varepsilon_{l-1} \sqrt{\sigma} \leq \mu_1 - \mu_2 < \varepsilon_{l-2} \) for \( l = 2 \) and case \( \varepsilon_{l-2} \leq \mu_1 - \mu_2 \leq \varepsilon_{l-2} \sqrt{\sigma} \) for \( l = 3 \) separately.

- \( \varepsilon_{l-1} \sigma \leq \mu_1 - \mu_2 < \varepsilon_{l-2} \) for \( l = 2 \): We know agents in level at least 2 will observe at least \( N_{FDP}^{T_1/2} \) pulls of arm \( a \) for \( a \in \{1, 2\} \). By \( W_2 \), for any agent in level at least 2, we have

\[
|\mu_a^l - \mu_a| \leq \sqrt{\frac{3 \log(T)}{\sigma N_{FDP}^{T_1/2}}}
\]

By Assumption 1, we have

\[
\hat{\mu}_1^l - \hat{\mu}_2^l \geq \hat{\mu}_1^l - \hat{\mu}_2^l - \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}} - \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}}
\]

\[
\geq \mu_1 - \mu_2 - \frac{3 \log(T)}{\sigma N_{FDP}^{T_1/2}} - \frac{3 \log(T)}{\sigma N_{FDP}^{T_1/2}} - \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}} - \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}}
\]

\[
\geq \frac{\sqrt{\sigma}}{4 \sqrt{N_{FDP}^{T_1}}} - \frac{\sqrt{\sigma}}{4 \sqrt{N_{FDP}^{T_1}}} - \frac{3 \log(T)}{\sigma N_{FDP}^{T_1/2}} - \frac{3 \log(T)}{\sigma N_{FDP}^{T_1/2}}
\]

\[
- \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}} - \frac{C_{\text{est}}}{\sqrt{\sigma N_{FDP}^{T_1/2}}}
\]

\[
> 0.
\]

Therefore agents in level at least 2 will all pull arm 1.

- \( \varepsilon_{l-1} \sigma \leq \mu_1 - \mu_2 < \varepsilon_{l-2} \) for \( l > 2 \): By claim 26, for any agent \( t \) in level at least \( l \), that agent will observe at least \( T_{l-1} \) arm \( a \) pulls. By \( W_2 \), we have

\[
|\mu_a^t - \mu_a| \leq \sqrt{\frac{3 \log(T)}{T_{l-1}}}
\]

27
By Assumption 1, we have
\[
\hat{\mu}_1^t - \hat{\mu}_2^t \geq \mu_1^t - \mu_2^t - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
\geq \mu_1 - \mu_2 - 2\sqrt{\frac{3 \log(T)}{T_{l-1}}} - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
\geq \sqrt{\frac{\sigma}{16T_{l-1}}} - 2\sqrt{\frac{3 \log(T)}{T_{l-1}}} - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
> 0.
\]

Therefore agents in level at least \( l \) will all pull arm 1.

\( \varepsilon_{l-2} < \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma \) for \( l = 3 \) (i.e. \( \varepsilon_1 < \mu_1 - \mu_2 < \varepsilon_1\sigma \)): By Claim 26, for any agent \( t \) in level at least 3, that agent will observe at least \( T_1N_{K,a}^2/2 \) arm \( a \) pulls (just from the first level). By \( W_2 \), we have
\[
|\bar{\mu}_a - \mu_a| \leq \sqrt{\frac{3 \log(T)}{\sigma^2 N_{FDP}^2 T_1/2}}.
\]

By Assumption 1, we have
\[
\hat{\mu}_1^t - \hat{\mu}_2^t \geq \mu_1^t - \mu_2^t - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
\geq \mu_1 - \mu_2 - 2\sqrt{\frac{3 \log(T)}{T_{l-1}}} - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
\geq \sqrt{\frac{\sigma}{16T_{l-1}}} - 2\sqrt{\frac{3 \log(T)}{T_{l-1}}} - \frac{2C_{\text{est}}}{\sqrt{T_{l-1}}}
\]
\[
> 0.
\]

Therefore agents in level at least 3 will all pull arm 1.

\( \varepsilon_{l-2} < \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma \) for \( l > 3 \): Since \( \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma < \varepsilon_{l-3} \), by Claim 26, for any agent \( t \) in level at least \( l \), that agent will observe at least \( T_{l-2}\sigma^2 \) arm \( a \) pulls (just from level \( l - 2 \)). By \( W_2 \), we have
\[
|\bar{\mu}_a - \mu_a| \leq \sqrt{\frac{3 \log(T)}{\sigma^2 T_{l-2}}}.
\]
By Assumption 1, we have
\[
\hat{\mu}_1^t - \hat{\mu}_2^t \geq \mu_1^t - \mu_2^t - \frac{2C_{est}}{\sqrt{\sigma^2 T_{l-2}}},
\]
\[
\geq \mu_1 - \mu_2 - 2\sqrt{\frac{3\log(T)}{\sigma^2 T_{l-2}}} - \frac{2C_{est}}{\sqrt{\sigma^2 T_{l-2}}},
\]
\[
\geq \frac{1}{4\sqrt{\sigma T_{l-2}}} - 2\sqrt{\frac{3\log(T)}{T_{l-1}}} - \frac{2C_{est}}{\sqrt{T_{l-1}}},
\]
\[
> 0.
\]
Therefore agents in level at least \( l \) will all pull arm 1.

Now we set the group sizes \( T_l \)'s as following. For \( l < L \),
\[
T_l = T^{L-1} \frac{T_{L-1}^L}{T^{L-1} + T_{L-2}^L + \cdots + T_1^L} / \sigma^3,
\]
and
\[
T_L = (T - T_1 \cdot L_K^{FDP} \cdot \sigma^2 - (T_2 + \cdots + T_{L-1}) \sigma^3) / \sigma^3.
\]
We restrict \( L \) to be at most \( \log(\ln(T) / \log(\sigma^4)) \) so that \( T_l/T_{l-1} \geq T^{1/2^l} \geq \sigma^4 \) for \( l = 2, \ldots, L-1 \). \( T_L \) is a little bit different because we want total number of agents to be \( T \).

By Claim 27, we know that the regret conditioned the intersection of clean events is at most
\[
\max \left( T_l L_K^{FDP} \sigma^2, \max_{l \geq 2} \varepsilon_{l-1} \sigma (T_l L_K^{FDP} \sigma^2 + T_2 \sigma^3 + \cdots + T_l \sigma^3) \right)
\]
\[
\leq \max \left( T_l L_K^{FDP} \sigma^2, \max_{l \geq 2} 2\varepsilon_{l-1} T_l \sigma^4 \right)
\]
\[
= O \left( T^{2^{L-1}/(2^{L-1})} \log^2(T) \right).
\]

Now we are going to change the parameters of the \( L \)-level recommendation policy a little bit and prove the below corollary. We will keep \( \sigma \) the same (i.e. \( \sigma = 2^{10} \log(T) \)). We are going to change \( L \) and \( T_1, \ldots, T_L \). We set \( L = \log(T) / \log(\sigma^4) \), \( T_l = (\sigma^4)^l \) for \( l = 1, \ldots, L-1 \) and \( T_L = (T - T_1 L_K^{FDP} \sigma^2 - \sigma^3 \sum_{l=2}^{L-1} T_l) / \sigma^3 \).

**Corollary 28** With the proper setting of \( L \) and \( T_1, \ldots, T_L \) described above, the \( L \)-level recommendation policy gets regret \( O(\min(1/\Delta, T^{1/2})\text{polylog}(T)) \). Here \( \Delta = |\mu_1 - \mu_2| \) and the \( L \)-level recommendation policy does not need to know \( \Delta \). Moreover, agent \( t \) observes a subhistory of size at least \( \Omega([t/\text{polylog}(T)]) \).

**Proof** Notice that in the proof of Theorem 25, by the end of Claim 27, the only constraint we need about \( T_l \)'s is that \( T_l/T_{l-1} \geq \sigma^4 \) for \( l = 2, \ldots, L-1 \) and \( T_1 \geq \sigma^4 \). And our new settings of \( T_l \)'s still satisfy this constraint. So we can reuse the proof of Theorem 25 till the end of Claim 27.

Recall in the proof of Theorem 25, \( \varepsilon_l = \Theta(1/\sqrt{T_l}) \) for \( l \in [L-1] \) and \( \varepsilon_0 = 1 \). Consider two cases:
\( \Delta < \varepsilon_{L-1}\sigma \). In this case, notice that even always picking the sub-optimal arm gives expected regret at most \( T(\mu_1 - \mu_2) = T\Delta = O(T^{1/2}\text{polylog}(T)) \). On the other hand, \( T^{1/2} = O(\text{polylog}(T)/\Delta) \). Therefore, the regret is \( O(\min(1/\Delta, T^{1/2})\text{polylog}(T)) \).

- \( \Delta \geq \varepsilon_{L-1}\sigma \). In this case, we can find \( l \in \{2, \ldots, L\} \) such that \( \varepsilon_{l-1}\sigma \leq \Delta < \varepsilon_{l-2}\sigma \). By Claim 27, we can upper bound the regret by

\[
\Delta \cdot (T_1 L_{K,FP}^{\text{PDP}} \sigma^2 + T_2 \sigma^3 + \cdots T_{l-1} \sigma^3)
= O(\Delta T_{l-1} \sigma^3)
= O(\Delta T_{l-2} \sigma^7)
= O(\Delta \cdot \frac{1}{\varepsilon_{l-2}^2} \cdot \sigma^6)
= O(\Delta \cdot \frac{1}{\Delta^2} \cdot \sigma^8)
= O(\text{polylog}(T)/\Delta).
\]

We also have \( 1/\Delta \leq 1/(\varepsilon_{L-1}\sigma) = O(T^{1/2}) \). Therefore, the regret is \( O(\min(1/\Delta, T^{1/2})\text{polylog}(T)) \).

Finally we discuss about the subhistory sizes. We know that agents in level \( l \) observes the history of all agents below level \( l - 2 \) (including level \( l - 2 \)). It is easy to check that the ratio between the number of agents below level \( l \) and the number of agents below level \( l - 2 \) is bounded by \( O(\text{polylog}(T)) \). Therefore our statement about the subhistory sizes holds.

Here we discuss about how to extend Theorem 25 and Corollary 28 to the case when \( K \) is a constant larger than 2. As the proof is very similar to the proofs of Theorem 25 and Corollary 28, we only provide a proof sketch of what changes to make.

**Theorem 29** Theorem 25 and Corollary 28 can be extended to the case when \( K \) is constant larger than 2. In the extension of Corollary 28, \( \Delta \) is defined as the difference between means of the best and the second best arm.

**Proof Sketch.** We still wlog assume arm 1 has the highest mean (i.e. \( \mu_1 \geq \mu_a, \forall a \in \mathcal{A} \)). We first extend the clean events (i.e. \( W_1, W_2, W_3, W_4 \)) in Theorem 25 to the case when \( K \) is larger than 2. \( W_1 \) and \( W_2 \) extend naturally: we still set \( W_1 = \bigcap_{a,s} W_1^{a,s} \) and \( W_2 = \bigcap_{t,a,t_1,t_2} W_2^{t,a,t_1,t_2} \).

The difference is that now \( a \) is taken over \( K \) arms instead of 2 arms. For \( W_3 \), we change the definition \( W_3^{l,a} = \bigcup_a \left( W_3^{l,a,a,\text{high}} \cap \left( \bigcap_{a' \neq a} W_3^{l,a',a,\text{low}} \right) \right) \) and \( W_3 = \bigcap_{l,a} W_3^{l,a} \). We extend \( W_4 \) in a similar way: define \( W_4^{a} \) as \( \bigcup_a \left( W_4^{a,a,\text{high}} \cap \left( \bigcap_{a' \neq a} W_4^{a',a,\text{low}} \right) \right) \) and \( W_4 = \bigcap_{a} W_4^{a} \). Since \( K \) is a constant, it’s easy to check that the same proof technique shows that the intersection of these clean events happen with probability \( 1 - O(1/T) \). So the case when some clean event does not happen contributes \( O(1) \) to the regret.

Now we proceed to extend Claim 26 and Claim 27. The statement of Claim 26 should be changed to “For any arm \( a \) and \( 2 \leq l \leq L \), if \( \mu_1 - \mu_a \leq \varepsilon_{l-1} \), then for any \( u \in [\sigma] \), there are at least \( T_l \) pulls of arm \( a \) in groups \( G_{l,u,1}, G_{l,u,2}, \ldots, G_{l,u,\sigma} \) and there are at least \( T_l \sigma (\sigma - 1) \) pulls of arm \( a \) in the \( l \)-th level \( \Gamma \)-groups”. The statement of Claim 27 should be changed to “For any \( 2 \leq l \leq L \), if \( \varepsilon_{l-1} \sigma \leq \mu_1 - \mu_a < \varepsilon_{l-2} \sigma \), there are no pulls of arm \( a \) in groups with level \( l, \ldots, L \).”
The proof of Claim 27 can be easily changed to prove the new version by changing “arm 2” to “arm a”. The proof of Claim 26 needs some additional argument. In the proof of Claim 26, we show that \( \hat{\mu}_t^a - \hat{\mu}_t^{3-a} > 0 \) for agent \( t \) in the chosen groups. When extending to more than 2 arms, we need to show \( \hat{\mu}_t^a - \hat{\mu}_t^{a'} > 0 \) for all arm \( a' \neq a \). The proof of Claim 26 goes through if \( \mu_1 - \mu_{a'} \leq \varepsilon_{l-2} \) since then there will be enough arm \( a' \) pulls in level \( l-1 \). We need some additional argument for the case when \( \mu_1 - \mu_{a'} > \varepsilon_{l-2} \). Since \( \mu_1 - \mu_{a'} > \varepsilon_{l-2} > \varepsilon_{l-1} \sigma \), we can use the same proof of Claim 27 (which rely on Claim 26 but for smaller \( l \)'s) to show that there are no arm \( a' \) pulls in level \( l \) and therefore \( \hat{\mu}_t^a - \hat{\mu}_t^{a'} > 0 \).

Finally we proceed to bound the regret conditioned on the intersection of clean events happens. The proofs of Theorem 25 and Corollary 28 bound it by consider the regret from pulling the suboptimal arm (i.e. arm 2). When extending to more than 2 arms, we can do the exactly same argument for all arms except arm 1. This will blow up the regret by a factor of \( (K - 1) \) which is a constant.