# Incentivizing Exploration with Selective Data Disclosure*

Nicole Immorlica † Jieming Mao ‡ Aleksandrs Slivkins §
Zhiwei Steven Wu ¶

**Abstract**

We study the design of rating systems that incentivize efficient social learning. Agents arrive sequentially and choose actions, each of which yields a reward drawn from an unknown distribution. A policy maps the rewards of previously-chosen actions to messages for arriving agents. The *regret* of a policy is the difference, over all rounds, between the expected reward of the best action and the reward induced by the policy. Prior work proposes policies that recommend a single action to each agent, obtaining optimal regret under standard rationality assumptions. We instead assume a frequentist behavioral model and, accordingly, restrict attention to *disclosure policies* that use messages consisting of the actions and rewards from a subsequence of past agents, chosen ex ante. We design a policy with optimal regret in the worst case over reward distributions. Our research suggests three components of effective polices: independent focus groups, group aggregators, and interlaced information structures.

*A preliminary abstract appeared at the *ACM Conf. on Economics and Computation (ACM EC)*, 2020. A working paper (with all technical results except robustness) has been available at https://arxiv.org/abs/1811.06026 since Nov 2018.

†Microsoft Research, New York, NY. Email: nicimm@microsoft.com.

‡Google, New York, NY. Email: maojm@google.com.
Research done during an internship at MSR-NYC.

§Microsoft Research, New York, NY. Email: slivkins@microsoft.com.

¶Carnegie-Mellon University, Pittsburgh, PA. Email: zstevenwu@cmu.edu.
Research done during a postdoc at MSR-NYC.

# Contents

**1 Introduction** **1**

**2 Related work** **5**

**3 Our model** **8**
3.1 Discussion: conceptual aspects . . . . . 11
3.2 Discussion: technical aspects . . . . . 13
3.3 Connection to multi-armed bandits . . . . . 14

**4 A simple two-level policy** **15**

**5 Adaptive exploration with a three-level policy** **19**

**6 Optimal regret with a multi-level policy** **23**

**7 Robustness** **27**

**8 Detailed comparison to prior work on incentivized exploration** **29**

**9 Conclusions** **31**

**Bibliography** **32**

**Appendix A Proofs** **37**
A.1 Preliminaries . . . . . 37
A.2 The two-level policy: proof of Theorem 4.3 . . . . . 37
A.3 The "global" counterexample: proof for Example 4.8 . . . . . 39
A.4 The three-level policy: proof of Theorem 5.2 . . . . . 39
A.5 The multi-level policy . . . . . 44

# 1 Introduction

A prominent feature of online platform markets is the pervasiveness of reviews and ratings. Unlike its brick-and-mortar competitors, Amazon accompanies its products by hundreds if not thousands of reviews and ratings from past consumers. Companies like Yelp and TripAdvisor have built entire business models on the premise of providing users with crowdsourced information about dining and hotel options so that they may make more informed choices.

The review and rating ecosystem creates a deep dilemma for online market designers. On the one hand, platforms would like to allow each consumer to make an informed choice by presenting the most comprehensive and comprehensible information. On the other hand, platforms need to encourage consumers to *explore* infrequently-selected alternatives in order to learn more about them. Extensive exploration may be required in settings, like ours, where the reward of an alternative is stochastic. The said exploration, while beneficial for the common good, is often misaligned with incentives of individual consumers. Being short-lived, individuals prefer to *exploit* available information, selecting alternatives that look best based on this information. This behavior can cause *herding* in which all agents take a sub-optimal alternative if, for example, agents see all prior ratings. Aside from such extreme behaviors, some alternatives may get explored at a very suboptimal rate, or suffer from selection bias.

Prior work leveraged information asymmetry to mitigate this tension between exploration and exploitation. The platform chooses a single recommendation for each consumer based on past ratings. Assuming, as is standard, that consumers are Bayesian rational and the platform has the power to commit, platforms can incentivize sufficient exploration to enable social learning asymptotically. However, these assumptions can be problematic in practice: consumers may hesitate to follow recommendations because of limited rationality, a preference for detailed and interpretable information, or insufficient trust in the platform's commitment power.

Our work also leverages information asymmetry to induce social learning, but does so in a behavioral model and with a restricted class of platform policies. We restrict the platform to *order-based* disclosure policies which provide each consumer with a subhistory of past ratings. Specifically, a partial order on the arrivals

is fixed ex ante (and can be made public w.l.o.g.), and each consumer observes full history for everyone who precedes her in this partial order. Put differently, an order-based disclosure policy constructs a communication network for the consumers, and lets them engage in social learning on this network. We assume consumers act like *frequentists*: they use the empirical mean of past ratings to estimate the rewards of alternatives. This is justified because each provided subhistory is *unbiased*: cannot be biased to make a particular action look good, and *transitive*: contains the information sets of all consumers therein. The latter property (and the assumption that consumers do not receive idiosyncratic signals) relieves a consumer of the need to reason about the rationale for the observed prior choices.

Our framework provides several key benefits. Our behavioral model only needs to define how consumers interact with the *full-disclosure policy* which reports the entire history. This is because, due to the unbiasedness and transitivity properties mentioned above, the only ratings that can possibly influence a consumer's beliefs are those included in her subhistory. Beliefs that follow empirical means are arguably quite reasonable for such settings, especially when databases are large. Our frequentist behavioral model captures this intuition, allowing a consumer to choose from a wide range of alternatives consistent with her confidence intervals. Moreover, order-based disclosure policies are arguably easier to audit than the complex code bases of general policies, thereby weakening the need for commitment. In contrast to prior work, our framework therefore relaxes rationality and commitment assumptions without abusing them, and provides detailed interpretable information to consumers.

We design several order-based disclosure policies in the context of this framework, of increasing complexity and improving performance guarantees. Our policies intertwine subhistories in a certain way, provably providing consumers with enough information to converge on the optimal alternative. Our best policy matches the best possible convergence rates, even absent incentive constraints. This policy also ensures that each consumer sees a substantial fraction of the prior history.

Our work suggests the importance of several design considerations. First, we observe that independent *focus groups* provide natural exploration due to random fluctuations in observed rewards. These natural experiments can then be provided to future agents to enable them to make optimal decisions. Second, we observe that improved learning rates require *adaptive exploration* which gradually zooms in on

the better alternatives. For example, if the focus groups learn the optimal alternative quickly, then this information should be propagated; otherwise additional exploration is required. This adaptivity can be achieved, even with subhistories chosen ex ante, by introducing *group aggregators* that see the subhistory of some, but not all, focus groups. Third, optimal learning rates require the *reuse of observations*; otherwise too many consumers make choices with limited information. The reused observations must be carefully interlaced to avoid contamination between experiments.

We start with a simple policy which runs the full-disclosure policy "in parallel" on several disjoint subsets of consumers (the "focus groups" mentioned above), collects all data from these runs, and discloses it to all remaining consumers. We think of this policy as having two "levels": Level 1 contains the parallel runs, and Level 2 is everyone else (corresponding to exploration and exploitation, respectively). While this policy provably avoids herding on a suboptimal alternative, it is still inefficient because it either over-explores bad alternatives, or under-explores the good-but-suboptimal ones.

Our next step is a proof-of-concept implementation of adaptive exploration. We focus on the case of two alternatives, and upgrade the simple two-level policy with a middle level. The consumers in this new level receive the data collected in some (but not all) parallel runs from the first level. These consumers explore only if the gap between the best and second-best alternative is sufficiently small, and exploit otherwise. When the gap is small, the runs in the first level do not have sufficient time to distinguish the two alternatives before herding on one of them. However, for each of these alternatives, there is some chance that it has an empirical mean reward significantly above its actual mean while the other alternatives have empirical mean reward significantly below its actual mean in any given first-level run. The middle-level consumers observing such runs will be induced to further explore that alternatives, collecting enough samples for the third-level consumers to distinguish the two alternatives.

The main result extends this construction to multiple "levels", connected in fairly intricate ways, using "group aggregators" and reusing information as discussed above. For each piece of our construction, we prove that agents' collective self-interested behavior guarantees a certain additional amount of exploration if, and only if, more exploration is needed at this point. The guarantee substantially

depends on the parameters of the problem instance (and on the level at which this piece resides), and critically relies on how the pieces are wired together.

Our framework is directly linked to *multi-armed bandits*, a popular abstraction for designing algorithms to balance exploration and exploitation. An order-based policy incentivizes agents to implement a multi-armed bandit algorithm, and agents' welfare is precisely the total reward of this algorithm. The two-level policy implements a well-known multi-armed bandit algorithm called *explore-then-exploit*, which explores in a pre-defined way for a pre-set number of rounds, then picks one alternative for exploitation and stays with it for the remaining rounds. Our multi-level policy implements a multi-armed bandit algorithm which can change its exploration schedule only a small number of times, each change-point corresponding to a level in our construction. (This is "adaptive exploration" with severely limited adaptivity, and not one of the standard bandit algorithms.)

We analyze our policies in terms of *regret*, a standard notion from the literature on multi-armed bandits, defined as the difference in total expected rewards between the best alternative and the algorithm.[1] We obtain sublinear regret rates, implying that the average expected reward converges to that of the best alternative. The multi-level policy matches the optimal regret rates for bandits, and hence learns at an optimal rate, for a constant number of alternatives. The two-level policy matches the standard (and very suboptimal) regret rates of bandit algorithms such as explore-then-exploit that do not use adaptive exploration. And the three-level policy admits an intermediate guarantee.

Our performance guarantees are robust in that they hold in the worst case over a class of reward distributions, and do not rely on priors. Moreover, our constructions are robust to small amounts of misspecification. First, all parameters can be increased by at most a constant factor (and the two-level construction allows a much larger amount of tweaking). Second, we accommodate some "information leakage", *e.g.*, rounds that are observable by other focus groups.

**Map of the paper.** Related work is in Section 2, with some detailed comparisons deferred to Section 8. Our model is defined and discussed in Section 3. The next three sections present our results on, resp., two-, three-, and multi-level policies. Section 7 is on robustness of these policies. All proofs are deferred to the appendix.

[1] Essentially, this is how much one *regrets* not knowing the best arm in advance.

## 2 Related work

The problem of incentivizing exploration via information asymmetry was introduced in (Kremer et al., 2014; Che and Hörner, 2018), under Bayesian rationality and (implicit) power-to-commit assumptions. A version closest to ours, which corresponds to multi-armed bandits with i.i.d. rewards and a Bayesian prior, was mostly resolved in (Kremer et al., 2014; Mansour et al., 2020, 2016; Sellke and Slivkins, 2020). The technical results come in a variety of flavors, concerning regret rates (Kremer et al., 2014; Mansour et al., 2020; Sellke and Slivkins, 2020), a black-box reduction from arbitrary bandit algorithms to incentive-compatible ones (Mansour et al., 2020), Bayesian-optimal policies for special cases (Kremer et al., 2014; Cohen and Mansour, 2019), and policies for exploring all "explorable" actions (Mansour et al., 2016). Several extensions were considered: to contextual bandits (Mansour et al., 2020), to repeated games and misaligned incentives (Mansour et al., 2016), to heterogenous agents (Immorlica et al., 2019), and to models with unavoidable information leakage (Bahar et al., 2016, 2019). Related, but technically different models feature: time-discounted utilities (Bimpikis et al., 2018); monetary incentives (Frazier et al., 2014; Chen et al., 2018); continuous information flow and a continuum of agents (Che and Hörner, 2018); and coordination of costly "exploration decisions", separate from "payoff-generating decisions" (Kleinberg et al., 2016; Liang and Mu, 2018; Liang et al., 2018).

The full-disclosure policy implements the "greedy" (exploitation-only) bandit algorithm, and suffers from herding on a suboptimal alternative with a positive-constant probability. However, a recent line of work (Kannan et al., 2018; Bastani et al., 2020; Raghavan et al., 2018; Acemoglu et al., 2019) proves that full disclosure avoids herding and provably performs well for heterogenous agents, under strong assumptions on the structure of rewards and diversity of agent types.[2]

A detailed comparison to results and modeling assumptions in prior work on incentivized exploration and full disclosure can be found Section 8.

Incentivized exploration is closely related to two prominent subareas of theoretical economics: *information design* and *social learning*. Information design (Bergemann and Morris, 2019; Kamenica, 2019) studies the design of information

[2]Kannan et al. (2018); Bastani et al. (2020); Raghavan et al. (2018) are framed in terms of multi-armed bandits. In our terms, they consider heterogenous agents with public types.

disclosure policies and incentives that they create. In particular, a single round of incentivized exploration is a version of the *Bayesian Persuasion* game (Kamenica and Gentzkow, 2011), where the signal observed by the principal is distinct from, but correlated with, the unknown "state". A strand of subsequent literature investigates conditions under which the optimal disclosure policy has a simple structure (Ivanov, 2015; Kolotilin, 2018; Machina and Siniscalchi, 2014; Mensch, 2019), including conditions that make assumptions on the agent behavior that, in some cases, exclude Bayesian rationality (Nikzad, 2019; Patil and Salant, 2020). Patil and Salant (2020) in particular assume, as we do, that agents form beliefs based on samples from a distribution without considering the mechanism that produced those samples.

A large literature on *social learning* studies agents that learn over time in a shared environment, with no principal to coordinate them. A prominent topic is the presence or absence of herding phenomena. Models vary across several dimensions, to wit: how an agent acquires new information; which information is transmitted to others; what is the structure / properties of the communication network; whether agents are long-lived or only act once; whether they optimize rewards (via Bayesian rationality or frequentist behavior), or merely follow a rule-of-thumb. While our work can be interpreted as coordinating social learning, all prior work studies models of social learning that are very different from ours. Below we discuss three lines of work in social learning that are most relevant.

First, "sequential social learning" posits that agents observe private signals, but only the chosen actions of neighbors are observable in the future; see Golub and Sadler (2016) for a survey. The early work focuses on sequential learning over a complete communication network, starting from Banerjee (1992); Welch (1992); Bikhchandani et al. (1992), with a very general result in Smith and Sørensen (2000). Further work considers the impact of the network topology on sequential learning. Acemoglu et al. (2011) and Lobel and Sadler (2015) show in a perfect Bayesian equilibrium, learning happens asymptotically if neighborhoods are sufficiently expansive or independent, features echoed in our own constructions. Sparse network topologies are studied in Bogachan and Kariv (2004); Banerjee and Fudenberg (2004); Acemoglu et al. (2014). To contrast these models with ours, we emphasize that the social planner only needs to choose the best action given the previous agents' signals, *i.e.*, only needs to *exploit*, whereas in our model it faces the explore-

exploit tradeoff. Also, herding occurs due to restricted information flow between the agents, whereas in our model it happens even with full disclosure.

Another line of work, starting from DeGroot (1974), posits that agents use "naive", mechanical rules-of-thumb, *e.g.*, form beliefs based on naive averaging of observations.[3] In particular, even naive agents learn asymptotically so long as the network is not too imbalanced (*e.g.*, Golub and Jackson, 2010). Chandrasekhar et al. (2020) show experimentally that such a behavioral model is a good predictor of human behavior in some scenarios. Dasaratha and He (2019) show similar results for sequential social learning. Theoretically, Dasaratha and He (2020) use this model of naivety to study the question of how to design the social network in a sequential learning model so as to induce optimal learning rates. They observe that silo structures akin to our two-level policy improve learning rates.

Third, "strategic experimentation", starting from Bolton and Harris (1999); Keller et al. (2005), studies long-lived learning agents that observe both actions and rewards of one another; see Hörner and Skrzypacz (2017) for a survey. This is similar to our work in that the social planner also solves a version of multi-armed bandits. The main difference is that the agents are long-lived and engage in a complex repeated game where each player deploys an exploration policy but would prefer to free-ride on exploration by others. There are also important technical differences. Agents exactly optimize their Bayesian-expected utility (using the Markov Perfect Equilibrium as a solution concept), whereas we consider a flexible frequentist model. Also, the social-planner problem is a very *different* bandit problem, with Bayesian prior, time-discounting, "safe" arm that is completely known, and "risky" arm that follows a stochastic process.

Absent incentive constraints, the so-called *exploration-exploitation tradeoff* has received much attention over the past decades in the relatively simple abstraction known as *multi-armed bandits*. In this abstraction, the social planner repeatedly selects from a set of actions (a.k.a. *arms*), each of which has a payoff drawn from an unknown fixed distribution. Over time, the planner can trade off *exploitation*, in which she picks an action to maximize expected reward, with *exploration*, in which she takes potentially sub-optimal actions to learn more about their rewards.

[3]Our frequentist agents may behave similarly, albeit with more justification (because the subhistories they observe are unbiased and transitive). The original paper of DeGroot (1974) and much subsequent work study agents that act repeatedly, updating their beliefs over rounds.

By coordinating actions across time, the planner can guarantee an average reward which converges to that of the optimal action in hindsight.

The vast literature on multi-armed bandits is covered in many books and surveys. We refer readers to Bubeck and Cesa-Bianchi (2012); Slivkins (2019); Lattimore and Szepesvári (2020) for background on regret-minimizing formulations, to Gittins et al. (2011); Bergemann and Välimäki (2006) for Bayesian and Markovian formulations, and to Cesa-Bianchi and Lugosi (2006); Slivkins (2019) for connections to economics and game theory. The most relevant thread in this literature, starting from Lai and Robbins (1985); Auer et al. (2002a), studies *stochastic bandits*: a basic model with i.i.d. rewards and no auxiliary structure, which corresponds to the social-planner version of our model. This basic model has been extended in many directions, with a lot of work on each: *e.g.*, payoffs with a known structure, non-stationary environments, and auxiliary payoff-relevant signals.

Exploration-exploitation problems with incentives issues naturally arise in a variety of scenarios, such as dynamic pricing (*e.g.*, Kleinberg and Leighton, 2003; Besbes and Zeevi, 2009; Badanidiyuru et al., 2018), dynamic auctions (*e.g.*, Athey and Segal, 2013; Bergemann and Välimäki, 2010; Kakade et al., 2013), pay-per-click ad auctions (*e.g.*, Babaioff et al., 2014; Devanur and Kakade, 2009; Babaioff et al., 2015), and human computation (*e.g.*, Ho et al., 2016; Ghosh and Hummel, 2013; Singla and Krause, 2013). These scenarios are not directly relevant to ours, so comparions would not be informative. For a unified perspective on exploration with incentives, see Chapter 11.6 in Slivkins (2019).

# 3 Our model

We study the multi-armed bandit problem in a social learning context, in which a platform (*principal*) faces a sequence of $T$ myopic consumers (*agents*). There is a set $\mathcal{A}$ of possible actions (*arms*). At each round $t \in [T]$, a new agent $t$ arrives, receives a *message* $m_t$ from the principal, chooses an arm $a_t \in \mathcal{A}$, and collects a reward $r_t \in \{0,1\}$.[4] The reward from pulling an arm $a \in \mathcal{A}$ is drawn independently from Bernoulli distribution $\mathcal{D}_a$ with an unknown mean $\mu_a$. The problem instance is defined by (known) parameters $|\mathcal{A}|$, $T$ and (unknown) mean rewards $\mu_a : a \in \mathcal{A}$.

The information structure is as follows. Each agent $t$ does not observe anything

[4]Throughout, we denote $[T] = \{1, 2, \ldots, T\}$.

from the previous rounds, other than the message $m_t$. The chosen arm $a_t$ and reward $r_t$ are observed by the principal (which corresponds, *e.g.*, , to the consumer leaving a rating or review on the platform).

The message $m_t$ could be arbitrary, *e.g.*, a recommended action or, in our case, a subset of past reviews. The principal chooses messages according to a decision rule called the *messaging policy*.

We assume that mean rewards are bounded away from 0 and 1, to ensure sufficient entropy in rewards. For concreteness, we posit $\mu_a \in [\frac{1}{3}, \frac{2}{3}]$.

**Regret.** We are interested in minimizing *regret*, formally defined as

$$\mathrm{Reg}(T) = T \max_{a \in \mathcal{A}} \mu_a - \sum_{t \in [T]} \mathbb{E}[\mu_{a_t}]. \tag{1}$$

The expectation is over the chosen arms $a_t$, which depend on randomness in rewards, and possibly in the policy. Thus, regret is the difference, in terms of the total expected reward, between the principal's policy and the first-best policy which knows the mean rewards a priori.

Following the literature on multi-armed bandits, we focus on the dependence on $T$, the number of agents. Assuming regret is sublinear in $T$, the average expected reward converges to that of the best arm at rate $\mathrm{Reg}(T)/T$. We are mainly interested in robust upper bounds on regret that hold *in the worst case* over all (valid) mean rewards. This provides guarantees (even) for a principal that has no access to a prior or simply does not make use of one due to extreme risk aversion.

We are also interested in performance of a policy at a given round $t$, as measured by *instantaneous regret* $\max_{a \in \mathcal{A}} \mu_a - \mathbb{E}[\mu_{a_t}]$, also known as *simple regret*. Note that summing it up over all rounds $t \in [T]$ gives $\mathrm{Reg}(T)$.

We use standard asymptotic notation to characterize regret rates: $O(f(T))$ and $\Omega(f(T))$ mean, resp., at most and at least $f(T)$, up to constant factors, starting from large enough $T$. Similarly, $\widetilde{O}(f(T))$ notation suppresses $\mathrm{polylog}(T)$ factors. Throughout, we assume that the number of arms $K = |\mathcal{A}|$ is constant. However, we explicitly note the dependence on $K$ when appropriate, *e.g.*, we use $O_K(\cdot)$ notation to note that the "constant" in $O()$ can depend on $K$ (and nothing else).

**Unbiased subhistories.** The *subhistory* for a subset of rounds $S \subset [T]$ is

$$\mathcal{H}_S = \{ (s, a_s, r_s) : s \in S \}. \tag{2}$$

This corresponds, for example, to a subset of past reviews. $\mathcal{H}_{[t-1]}$ is called the *full history* at time $t$. The *outcome* for agent $t$ is the tuple $(t, a_t, r_t)$.

We focus on messaging policies where the message in each round $t$ is $m_t = \mathcal{H}_{S_t}$ for some subset $S_t \subset [t-1]$. We assume that the subset $S_t$ is chosen ahead of time, before round 1 (and therefore does not depend on the observations $\mathcal{H}_{t-1}$). Such a message is called *unbiased subhistory*; it means the platform can not bias the set of reviews it shows a consumer, *e.g.*, by selecting only those in which a particular arm has positive ratings. To define subsets $S_t$, we fix a partial order on the rounds, and define each $S_t$ as the set of all rounds that precede $t$ in the partial order. The resulting disclosure policy is called *order-based*.

Order-based disclosure policies are *transitive*, in the following sense:

$$t \in S_{t'} \Rightarrow S_t \subset S_{t'} \qquad \text{for all rounds } t, t' \in [T]. \tag{3}$$

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 $a_t$.

For convenience, we will represent an order-based policy 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 order-based 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 $a \in \mathcal{A}$, and chooses an arm with a maximal estimate.[5] A simple instantiation 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, $\hat{\mu}_{t,a} = \frac{1}{2}$.

We allow a much more permissive model, where agents can form arbitrary reward estimates as long as they lie within some "confidence range" of the sample average. Formally, the model is characterized by the following assumptions (which we make without further notice).

[5]To simplify proofs, ties between the reward estimates are broken according to some fixed, deterministic ordering over the arms. This is to rule out adversarial manipulation of the tie breaking, and to ensure that all agents with the same data choose the same arm.

**Assumption 3.1.** *Reward estimates are close to empirical averages. Let $N_{t,a}$ and $\bar{\mu}_{t,a}$ denote the number of pulls and the empirical mean reward of arm $a$ in subhistory $m_t$. Then for some absolute constant $N_{\texttt{est}} \in \mathbb{N}$ and $C_{\texttt{est}} = \frac{1}{16}$, and for all agents $t \in [T]$ and arms $a \in \mathcal{A}$ it holds that*

$$\text{if } N_{t,a} \geq N_{\texttt{est}} \quad \text{then} \quad \left|\hat{\mu}_a^t - \bar{\mu}_a^t\right| < \frac{C_{\texttt{est}}}{\sqrt{N_{t,a}}} \tag{4}$$

$$\text{if } N_{t,a} = 0 \quad \text{then} \quad \hat{\mu}_a^t \geq 1/3. \tag{5}$$

*(NB: we make no assumption if $1 \leq N_{t,a} < N_{\texttt{est}}$.)*

The 1/3 threshold in Eq. (5) can be replaced with an arbitrary strictly positive constant, with very minor changes in the proofs. We just need to assume that the initial estimates are bounded away from zero. Some other alternative model variants are discussed in Section 3.2.

**Assumption 3.2.** *In each round $t$, the estimates $(\hat{\mu}_{t,a} : a \in \mathcal{A})$ depend only on the multiset $m'_t = \{ (a_s, r_s) : s \in S_t \}$, called* anonymized subhistory. *Each agent $t$ forms its estimates according to some function $f_t$ from anonymized subhistories to $[0,1]^{|\mathcal{A}|}$, so that $(\hat{\mu}_{t,a} : a \in \mathcal{A}) = f_t(m'_t)$. For each $t$, this function is drawn independently from some fixed (but otherwise arbitrary) distribution.*

Note that the estimators $(\hat{\mu}_{t,a} : a \in \mathcal{A})$ can be different for different arms, they can be randomized, and they can be arbitrarily correlated across arms. This allows, for instance, an agent to be optimistic about Chinese restuarants and pessimistic about Italian ones.

## 3.1 Discussion: conceptual aspects

We consider a model of incentivized exploration for which the "unrestricted version" — one with unrestricted rationality and commitment assumptions — is already well-studied, and focus on mitigating these assumptions. Several extensions are possible (*e.g.*, to heterogenous and/or long-lived agents, see Conclusions), but they are not well-understood even in the unrestricted version.

As mentioned in the Introduction, our model provides two key benefits:

(i) Effectively, we only need to make assumptions on how agents interact with the *full-disclosure policy*, rather than with an arbitrary messaging policy.

(ii) We allow a flexible frequentist choice model, whereby an agent is only required to be consistent with a (slightly narrower version of) her confidence intervals, and only after collecting $N_{\text{est}}$ samples of each arm.

Each agent observes full history for the relevant part of the mechanism, *as if* the full-disclosure policy were used. Indeed, by transitivity of the partial order, the only rounds that can possibly affect round $t$ are the ones in $S_t$. Rounds not in $S_t$ are as irrelevant to the agent arriving in round $t$ as anything else that happens outside the mechanism. Thus, as far as this agent is concerned, the relevant mechanism is one restricted to the rounds $S_t \cup \{t\}$, and the agent sees the full history thereof.

Put differently, our framework encourages agents to interpret the subhistory as a full set of data points collected by some policy. There is no reason to second-guess why a particular data point has been chosen (as neither the platform or the other agents can influence this choice), or what data has been seen by an agent when she chose her action (because all that data is included in the subhistory).

How would a frequentist agent choose an action given the full history of observations? She would construct a confidence interval on the expected reward of each action, taking into account the average reward of this action and the number of observations. The system can provide summary statistics, so that agents would not need to look at the raw data. Further, we allow agents to have strong initial beliefs, whose effect is eventually drowned out (due to $N_{\text{est}}$ in Assumption 3.1).

By virtue of having a frequentist choice model, we bypass a host of standard issues inherent in Bayesian choice models: we do not need to worry whether and to which extent the principal knows the prior, or whether users have correct beliefs, or whether they can handle the cognitive load of Bayesian reasoning. Moreover, we allow two deviations from rationality. First, we allow for a considerable amount of optimism or pessimism: an optimistic (resp., pessimistic) agent may estimate each action's expected reward as a value towards the top (resp., bottom) of its confidence interval. Second, we allow Softmax-like choices that randomize around the best actions. Indeed, each reward estimate can be randomized, as long as it falls in the corresponding confidence interval.

Nevertheless, our model *is* consistent with a version of Bayesian rationality. For example, suppose agents believe that rewards of each arm $a$ come from an independent Beta-Bernoulli prior, and the estimate $\hat{\mu}_a^t$ is the posterior mean reward given the subhistory $m_t$. Then the estimates satisfy Assumption 3.1 for a large

enough constant $N_{\mathtt{est}}$ which depends on the priors.[6] However, such beliefs would necessarily be *inconsistent* with our model of rewards, as they place positive probability outside of the $[1/3, 2/3]$ interval.[7]

While detail-oriented users may prefer to observe full data, our policies show all but a few past datapoints to all but a few users, and our main result shows a certain fraction of the full history to *all* users. Besides, even a small fraction of the full history would typically contain a large number of observations (preselected in an unbiased way), probably more than a typical user ever needs.

## 3.2 Discussion: technical aspects

Several technical aspects of our model are worth elaborating. First, while we focus on the paradigmatic case of Bernoulli rewards, we can handle arbitrary rewards $r_t \in [0,1]$ with only minor modifications to the analysis. It suffices to assume that the reward distribution for each arm places (at least) a positive-constant probability mass on, say, subintervals $[0, 1/4]$ and $[3/4, 1]$. Alternatively, we could round each reward $r_t$ as an independent Bernoulli draw with mean $r_t$, and only reveal these "rounded rewards" to the future agents instead of the true rewards, corresponding to a granular rating structure.

Second, our analysis relies on the assumption that the mean rewards lie in $[1/3, 2/3]$. This interval can be replaced with $[\epsilon, 1-\epsilon]$ for any fixed absolute constant $\epsilon > 0$ (which propagates throughout the analysis). The lower bound ensures that a sufficiently long string of low rewards of one arm can drive its reward estimate far below the mean reward any another arm. The upper bound ensures that the said string shows up with a reasonably large probability. Moreover, both bounds are needed to guarantee large variance in rewards, which we use to derive anti-concentration (via Berry-Esseen theorem).

Third, our model requires agents to form estimates of rewards that are below $1/3$ after observing a long sequence of low rewards. This is of course inconsistent with the assumption that $\mu_a \in [1/3, 2/3]$. We can remove this "unawareness assumption" and instead project all reward estimates into the $[1/3, 2/3]$ interval,[8] assuming

[6] This is because for Beta-Bernoulli priors the absolute difference between the posterior mean and the empirical mean scales as 1/#samples.

[7] This is needed for the posterior mean rewards to satisfy (4) even if the empirical mean reward falls below $1/3$, which in turn is necessary to get exploration going.

[8] That is, truncate the reward estimate at $1/3$ (resp., $2/3$) if it becomes too low (resp., too high).

random tie-breaking. This variant works with minimal changes. Alternatively, we could argue that frequentist agents are unaware of the restriction on mean rewards because they have incomplete information and/or are unsophisticated.

## 3.3 Connection to multi-armed bandits

Regret in our model can be directly compared to regret in the stochastic bandit problem with the same mean rewards. Following the literature, we define the *gap parameter* $\Delta$ as the difference between the largest and second largest mean rewards. The gap parameter is not known (to the principal in incentivized exploration, or to the algorithm in bandits); large $\Delta$ corresponds to "easy" problem instances. The literature is mainly concerned with asymptotic upper bounds on regret in terms of the time horizon $T$, as well as parameters $\Delta$ and the number of arms $K$.

Optimal regret rates are as follows (Auer et al., 2002a,b; Lai and Robbins, 1985):

$$\text{Reg}(T) \leq O\Big(\min\Big(\sqrt{KT\log T},\ \tfrac{K}{\Delta}\log T\Big)\Big). \tag{6}$$

This includes a *worst-case* regret rate $O(\sqrt{KT\log T})$ which applies to all problem instances, and a *gap-dependent* regret rate of $O(\frac{K}{\Delta}\log T)$. We match both regret rates for a constant number of arms. Either regret rate can only be achieved via *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 number $N$, then chooses one arm and sticks with it till the end. More generally, *exploration-separated* algorithms have a property that 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. Such algorithms suffer from $\Omega(T^{2/3})$ regret, both in the worst case and for each problem instance.[9]

[9]More precisely, exploration-separated algorithms exhibit a tradeoff between the worst-case and per-instance performance: if the algorithm achieves regret $O(T^\gamma)$ for all instances, for some $\gamma \in [2/3, 1)$, then its regret for each instance can be no better than $\Omega\big(T^{2(1-\gamma)}\big)$. The latter is $\Omega(T^{2/3})$ when $\gamma = 2/3$. All these lower bounds are from Babaioff et al. (2014). (They consider a closely related but technically different setting which can be easily "translated" into ours.) The worst-case lower bound has been "folklore knowledge" in the community long before that.

# 4 A simple two-level policy

We first design a simple policy that exhibits asymptotic learning (i.e., sublinear regret). While not achieving an optimal regret rate, this policy illuminates a key feature: initial agents are partitioned into *focus groups*. Each agent sees the history for all previous agents in the same focus group (and nothing else). The information generated by these focus groups is then presented to later agents. We think of this policy as having two *levels*: the exploration level containing the focus groups, followed by the exploitation level. All agents in the latter observe full history.

Although simple, the two-level policy does exhibit some subtleties. First, it is important that the focus groups are independent. For example, a few initial agents observable by all other agents may induce herding on a suboptimal arm; we flesh out this point in Example 4.8. Second, it is important that each focus group has a linear information flow. For example, the first few agents acting in isolation (and biased in favor of arm 1) may force high-probability herding within the focus group, preventing the natural exploration that we rely on; see Example 4.9. Third, it is important that there are enough focus groups and agents therein, but not too many. Indeed, we need enough agents in each focus group to overpower the initial biases (as expressed by the reward estimators given less than $N_{\texttt{est}}$ samples). Having enough focus groups ensures that the natural exploration succeeds. However, agents in the focus groups would have limited information and may make suboptimal choices, so having too many of them would induce high regret. Fourth, agents with limited observations are not very restricted by our assumptions. For example, they may be systematically pessimistic about the optimal arm, and optimistic about a suboptimal one, so as to consistently choose the latter. Yet, the focus groups provide enough data to the future agents to overcome these biases.

We first describe the structure of a single focus group. Consider a disclosure policy that reveals the full history in each round $t$, *i.e.*, $S_t = [t-1]$; we call it the *full-disclosure policy*. The info-graph for this policy is a simple path. Intuitively, all agents in a the path of this full-disclosure policy are in a single focus group. We use this policy as a "gadget" in our constructions, formulated as follows:

**Definition 4.1.** A subset of rounds $S \subset [T]$ 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 (at least) a positive-constant probability, a full-disclosure path of constant length suffices to sample each arm at least once. This happens due to stochastic variation in outcomes; some agents in a focus group will get uncharacteristically bad rewards from an arm, inducing others to pull a different arm. We will build on this fact throughout.

**Lemma 4.2.** *There exist numbers $L_K^{\text{FDP}} > 0$ and $p_K^{\text{FDP}} > 0$ that depend only on $K$, the number of arms, with the following property. Consider an arbitrary disclosure policy, and let $S \subset [T]$ be a full-disclosure path in its info-graph, of length $|S| \geq L_K^{\text{FDP}}$. Under Assumption 3.1, with probability at least $p_K^{\text{FDP}}$, it holds that subhistory $\mathcal{H}_S$ contains at least one sample of each arm $a$.*

*Proof.* Fix any arm $a$. Let $L_K^{\text{FDP}} = (K-1) \cdot N_{\texttt{est}} + 1$ and $p_K^{\text{FDP}} = (1/3)^{L_K^{\text{FDP}}}$. We will condition on the event that all the realized rewards in $L_K^{\text{FDP}}$ rounds are 0, which occurs with probability at least $p_K^{\text{FDP}}$ under Assumption 3.1. 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 $N_{\texttt{est}} + 1$ rounds. Let $t$ be the round in which arm $a'$ is pulled exactly $N_{\texttt{est}} + 1$ times. By Assumption 3.1, we know

$$\hat{\mu}_{a'}^t \leq 0 + C_{\texttt{est}}/\sqrt{N_{\texttt{est}}} \leq C_{\texttt{est}} < 1/3.$$

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 $a$. □

Now let us define the *two-level policy*: an order-based disclosure policy which follows the "explore-then-exploit" paradigm. The "exploration level" comprises the first $N = T_1 \cdot L_K^{\text{FDP}}$ rounds, and consists of $T_1$ full-disclosure paths of length $L_K^{\text{FDP}}$ each, where $T_1$ is a parameter (*i.e.*, $T_1$ independent focus groups). In the "exploitation level", each agent $t > N$ receives the full history, *i.e.*, $S_t = [t-1]$. [10] The info-graph for this disclosure policy is shown in Figure 1.

[10] For the regret bounds, it suffices if each agent in the exploitation level only observes the history from the exploration level, or any superset thereof.

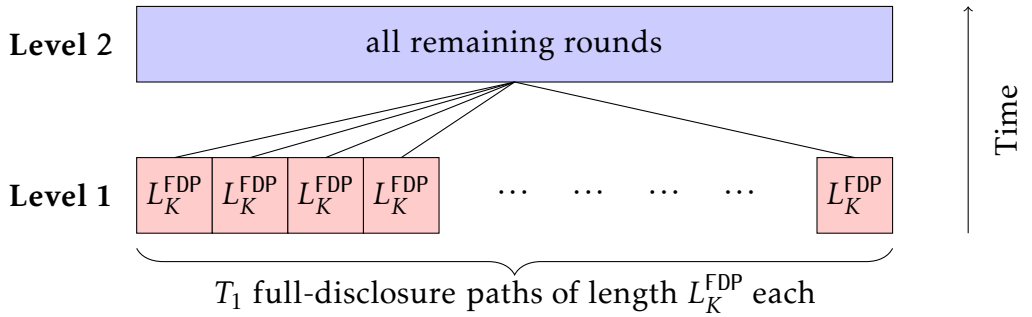


**Figure 1:** Info-graph for the 2-level policy.

We show that this policy 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 (at least) a positive-constant probability, using many full-disclosure paths "in parallel" ensures that sufficiently many samples of this arm are collected with very high probability.

**Theorem 4.3.** *The two-level policy with parameter* $T_1 = T^{2/3}(\log T)^{1/3}$ *achieves regret*

$$\text{Reg}(T) \leq O_K\left(T^{2/3}(\log T)^{1/3}\right).$$

*Remark* 4.4. All agents $t > T^{2/3}(\log T)^{1/3}$ (*i.e.*, all but the vanishingly small fraction of agents who are in the exploration level) observe full history, and pull an arm with instantaneous regret at most $\widetilde{O}\left(T^{-1/3}\right)$.

*Remark* 4.5. Each full-disclosure path can be made arbitrarily longer, and more full-disclosure paths can be added (of arbitrarily length), as long as the total number of Level-1 agents increases by at most a constant factor. Same regret bounds are attained with minimal changes in the analysis.

*Remark* 4.6. 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 proof can be found in Appendix A.2. One important quantity is the expected number of samples of a given arm $a$ collected by a full-disclosure path $S$ of length $L_K^{\text{FDP}}$ (*i.e.*, present in the subhistory $\mathcal{H}_S$. Indeed, this number, denoted $N_{K,a}^{\text{FDP}}$, is the same for all such paths. Then,

**Lemma 4.7.** *Suppose the info-graph contains $T_1$ full-disclosure paths of $L_K^{\mathsf{FDP}}$ rounds each. Let $N_a$ be the number of samples of arm $a$ collected by all paths. Then with probability at least $1-\delta$,*

$$\left|N_a - N_{K,a}^{\mathsf{FDP}} T_1\right| \leq L_K^{\mathsf{FDP}} \cdot \sqrt{T_1 \log(2K/\delta)/2} \quad \text{for all } a \in \mathcal{A}.$$

Let us flesh out the two counterexamples mentioned above. The first counterexample tweaks the "global" structure of the network. A few initial agents see the full history, causing them to herd on a suboptimal arm with constant probability. If all future agents then see the history of these initial agents, the inefficient arm may persist indefinitely. This might happen if, for example, the initial agents are celebrities, and their experiences leak to future agents outside the platform.

**Example 4.8** (Global)**.** Posit $K = 2$ arms such that $3/4 \geq \mu_1 > \mu_2 > 1/4$. Suppose Assumption 3.1 holds with $N_{\texttt{est}} = 2$ so that each arm is chosen in the first two rounds, and subsequently the mean reward of each arm $a$ is estimated by the sample average (*i.e.*, $\hat{\mu}_a^t := \bar{\mu}_a^t$ for all rounds $t > 2$). If each of the first $N$ rounds are observable by all subsequent agents, for a large enough $N = \Omega(\sqrt{\log(T)})$, then with (at least) a positive-constant probability it holds that all agents $t > N$ choose arm 2.

The proof can be found in Appendix A.3.

The second counterexample is "local" in nature, tweaking the information flow in a particular focus group so that the first few agents act in isolation from each other (and everyone else). This may happen, for example, if their reviews are submitted and/or processed with a substantial delay. Suppose these initial agents are pessimistic about arm 1, so that each one in isolation pulls arm 2. This builds certainty about the mean reward of arm 2 which, for an appropriate setting of parameters, may exceed the initial reward estimate for arm 1. Then later agents viewing all this information will fail to pull arm 1 with high probability.

**Example 4.9** (Local)**.** Suppose there are only two arms, all agents initially prefer arm 1, and have the same initial reward estimate $\hat{\mu}_2$ for arm 2. Consider a full-disclosure path $P$ starting at round $t_0$. Suppose agent $t_0$ observes $N$ "leaf agents" (each of which does not observe anybody else). Then, for any absolute constant $\mu_1 > \hat{\mu}_2$ and a sufficiently large $N = \Omega(\sqrt{\log(T)})$, each agent in $P$ will not try arm 2 with probability, say, at least $1 - O(T^{-2})$.

# 5 Adaptive exploration with a three-level policy

The two-level policy from the previous section implements the explore-then-exploit paradigm using a basic design with focus groups. The next challenge is to implement *adaptive exploration*, and go below the $T^{2/3}$ barrier. Standard multi-armed bandit algorithms achieve this by pulling sub-optimal arm on occasion, when and if the available information requires it. However, we can not adaptively add focus groups since we must fix our policy ahead of time.

Instead, we accomplish adaptive exploration using a construction that adds a middle level to the info-graph. Agents in this middle level are partitioned into subgroups, each responsible for aggregating information from a subset of focus groups; we call these agents *group aggregators*. For simplicity, we assume $K = 2$ arms. When one arm is much better than the other, group aggregators have enough information to discern it and *exploit*. However, when the two arms are close, group aggregators will be induced to pull different arms (depending on the outcomes in their particular focus groups), which induces additional exploration. This construction also provides intuition for the main result, the multi-level construction presented in the next section.

**Construction 5.1.** *The* three-level policy *is an order-based disclosure policy defined as follows. 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 $L_K^{\mathsf{FDP}}$ each, and each agent $t$ in the exploitation level sees full history (see Figure 2).* [11]

*The middle level consists of $\sigma$ disjoint subsets of $T_2$ agents each, called* second-level groups. *All nodes in a given second-level group $G$ are connected to the same nodes outside of $G$, but not to one another.*

*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 $T_1 \cdot \sigma \cdot L_K^{\mathsf{FDP}}$ 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*

[11] It suffices for the regret bounds if each agent in the exploitation level only observes the history from exploration (*i.e.*, from all agents in the first two levels), or any superset thereof.

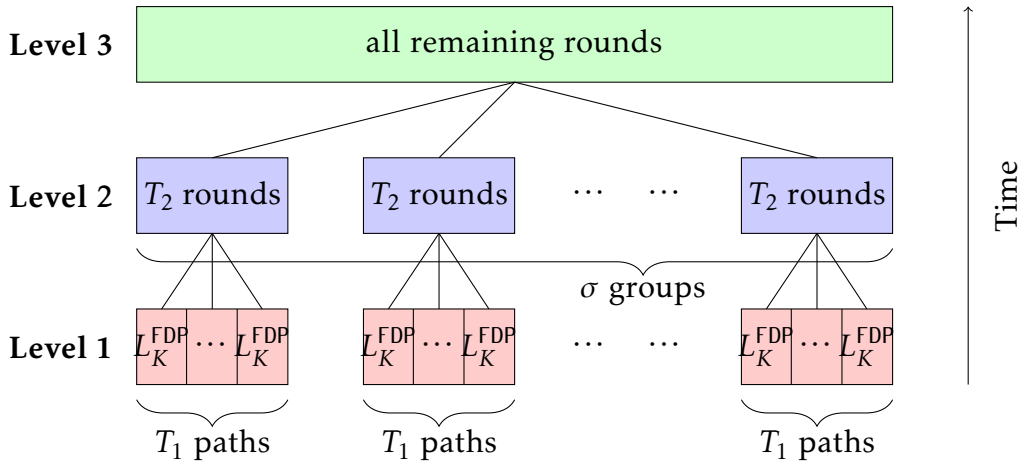


**Figure 2:** Info-graph for the three-level policy. Each red box in level 1 corresponds to $T_1$ full-disclosure paths of length $L_K^{\text{FDP}}$ each.

$\mathcal{H}_S$, *where* $S$ *is the set of all rounds in* $G$.

In more detail, the key idea is as follows. Consider the gap parameter $\Delta = |\mu_1 - \mu_2|$. If it is 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. Then 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 5.2.** *For two arms, the three-level policy achieves regret*

$$\text{Reg}(T) \leq O\Big(T^{4/7} \log T\Big).$$

*This holds for parameters* $T_1 = T^{4/7} \log^{-1/7}(T)$, $\sigma = 2^{10} \log(T)$, *and* $T_2 = T^{6/7} \log^{-5/7}(T)$.

*Remark* 5.3. All agents $t > \widetilde{O}(T^{6/7})$ (*i.e.*, all but a vanishingly small fraction of agents who are in the first two levels) observe full history, and pull an arm with instantaneous regret $\widetilde{O}\Big(T^{-3/7}\Big)$.

Let us sketch the proof; the full proof can be found in Appendix A.4.

**The "good events".** We establish four "good events" each of which occurs with high probability.

($\texttt{event}_1$) *Exploration in Level 1:* Every first-level group collects at least $\Omega(T_1)$ samples of each arm.

($\texttt{event}_2$) *Concentration in Level 1:* Within each first-level group, empirical mean rewards of each arm $a$ concentrate around $\mu_a$.

($\texttt{event}_3$) *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 the other arm look much worse than its actual mean.

($\texttt{event}_4$) *Concentration in prefix:* The empirical mean reward of each arm $a$ concentrates around $\mu_a$ in any prefix of its pulls. (This ensures 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 Appendix A.1). For example, $\texttt{event}_2$ 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, denote $\texttt{conf}(n) = \sqrt{\log(T)/n}$. In words, this is a confidence term, up to constant factors, for $n$ independent random samples.

The simplest case is very small gap, trivially yielding an upper bound on regret.

**Claim 5.4** (Negligible gap)**.** *If* $\Delta \leq 3\sqrt{2} \cdot \texttt{conf}(T_2)$ *then* $\text{Reg}(T) \leq O(T^{4/7}\log^{6/7}(T))$.

Another simple 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 $\texttt{event}_1$ and $\texttt{event}_2$.

**Lemma 5.5** (Large gap)**.** *If* $\Delta \geq 4\sum_{a\in\mathcal{A}} \texttt{conf}\left(N^{\text{FDP}}_{K,a} \cdot T_1\right)$ *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 *all* first-level groups, which enables them to correctly identify the best arm.

**Lemma 5.6** (Medium gap)**.** *All agents pull arm 1 in the third level, when* $\Delta$ *satisfies*

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

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 $\texttt{event}_3$, 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 3.1 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 5.7** (Small gap)**.** *All agents pull arm 1 in the third level, when*

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

**Wrapping up: proof of Theorem 5.2.** In negligible gap case, the stated regret bound holds regardless of what the policy 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 $\sigma \cdot L_K^{\mathsf{FDP}} \cdot T_1 = O(T^{4/7} \log T)$. In both intermediate cases, it suffices to bound the regret from the first and second levels, so

$$\mathrm{Reg}(T) \leq (\sigma\, T_1 \cdot L_K^{\mathsf{FDP}} + \sigma\, T_2) \cdot 4 \sum_{a \in \mathcal{A}} \operatorname{conf}\left(N_{K,a}^{\mathsf{FDP}} \cdot T_1\right) = O(T^{4/7} \log^{6/7}(T)).$$

Therefore, we obtain the stated regret bound in all cases.

# 6 Optimal regret with a multi-level policy

We extend our three-level policy to a more adaptive multi-level policy in order to achieve the optimal regret rate of $\widetilde{O}_K(\sqrt{T})$. This requires us to distinguish finer and finer gaps between the best and second-best arm. A naive approach would be to recursively apply the 2-level structure, creating a tree of group aggregators, each level responsible for successively larger information sets. This mimics the hierarchical information structure in many organizations, but it suffers large regret because the number of agents in focus groups grows exponentially. Furthermore, each of these agents is forced to make decisions with access to a vanishingly-small amount of history, which is undesirable in-and-of itself. In this section, we describe a method of interlacing information to reuse it without suffering from introduced correlations. This careful reuse of information is the third and final step in our journey towards policies with optimal learning rates.

On a very high level, our multi-level policy implement the limited-adaptivity framework for multi-armed bandits (Perchet et al., 2016), defined is as follows. Suppose a bandit algorithm outputs a distribution $p_t$ over arms in each round $t$, and the arm $a_t$ is then drawn 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. Optimal regret rate requires at least $O(\log\log T)$ adaptivity rounds, where each "level" $\ell \geq 2$ in our construction implements one adaptivity round. The limited-adaptivity bandit algorithm from Perchet et al. (2016) is much simpler compared to our construction below, as it can ensure the desired amount of exploration directly by choosing the appropriate alternatives.

We provide two results (for two different parameterizations of the same policy). The first result analyzes the $L$-level policy for an arbitrary $L \leq O(\log\log T)$, and achieves the root-$T$ regret rate with $O(\log\log T)$ levels.

**Theorem 6.1.** *There exists $L_{\max} = \Theta(\log\log T)$ such that for each $L \in \{3, 4, \ldots, L_{\max}\}$ there exists an order-based disclosure policy with $L$ levels and regret*

$$\mathrm{Reg}(T) \leq O_K\left(T^{\gamma} \cdot \mathrm{polylog}(T)\right), \quad \text{where } \gamma = \frac{2^{L-1}}{2^L - 1}.$$

*In particular, we obtain regret $O_K(T^{1/2}\,\mathrm{polylog}(T))$ with $L = O(\log\log(T))$.*

Our second policy achieves a gap-dependent regret guarantee, as per (6). This

policy has the same info-graph structure as the first one in Theorem 6.1, 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 6.1 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 6.2.** *There exists an order-based disclosure policy with $L = O(\log(T)/\log\log(T))$ levels such that for every bandit instance with gap parameter $\Delta$, the policy has regret*

$$\mathrm{Reg}(T) \leq O_K\Big(\min\Big(1/\Delta,\ T^{1/2}\Big)\cdot \mathrm{polylog}(T)\Big).$$

*Under this policy, each agent $t$ observes a subhistory of size at least $\Omega(t/\mathrm{polylog}(T))$.*

Note for constant number of arms, this result matches the optimal regret rate (given in Equation (6)) for stochastic bandits, up to logarithmic factors.

*Remark* 6.3. The multi-level policy can be applied to the first $T/\eta(T)$ agents only, for any fixed $\eta(T) = \mathrm{polylog}(T)$ (*i.e.*, , with reduced time horizon $T/\eta(T)$). Then the subsequent agents – which comprise all but $1/\eta(T)$-fraction of the agents – can observe the full history and enjoy instantaneous regret $\widetilde{O}\big(T^{-1/2}\big)$. The regret bounds from both theorems carry over. This extension requires only minimal modifications to the analysis, which are omitted.

Let us present the main techniques in our solution, focusing on the case of $K = 2$ arms; the full proofs are deferred to Section A.5.

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. In particular, each level will be responsible for some range of the gap parameter, collecting enough samples to rule out the bad arm if the gap parameter falls in this range. However, we need to introduce two main modifications in the info-graph to accommodate some new challenges.

**Interlacing connections between levels.** A tempting approach, described intuitively at the beginning of this section, generalizes the three-level policy to build an $L$-level info-graph with the structure of a $\sigma$-ary tree: for every $\ell \in \{2,\ldots,L\}$, each $\ell$-level group observes the sub-history from a disjoint set $\sigma$ groups in level $(\ell-1)$.

The disjoint sub-histories observed by all the groups in level $\ell$ are independent, and under the small gap regime (similar to Lemma 5.7) it ensures that each arm $a$ has a "lucky" $\ell$-level group of agents that only pull $a$. This "lucky" property is crucial for ensuring that both arms will be explored in level $\ell$.

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 6.2. 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 $\Theta(\log T)$ $\ell$-th level groups that observe disjoint sub-histories that take place in level $(\ell - 1)$. Moreover, as long as the group size in levels lower than $(\ell - 1)$ are substantially smaller than group size of level $\ell - 1$, the "lucky" property does not break even if different groups in level $\ell$ observe overlapping sub-history from levels $\{1, \ldots, \ell - 2\}$.

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 $\ell$-th level are labeled as $G_{\ell,u,v}$ for $u, v \in [\sigma]$. For any $\ell \in \{2, \ldots, L\}$ and $u, v, w \in [\sigma]$, agents in group $G_{\ell,u,v}$ see the history of agents in group $G_{\ell-1,v,w}$ (and by transitivity all agents in levels below $\ell - 1$). See Figure 3 for a visualization of simple case with $\sigma = 2$). Two observations are in order:

(i) Consider level $(\ell-1)$ and fix the last group index to be $v$, and consider the set of groups $\mathcal{G}_{\ell-1,v} = \{G_{\ell-1,i,v} \mid i \in [\sigma]\}$ (e.g. $G_{\ell-1,1,1}$ and $G_{\ell-1,2,1}$ circled in red in the Figure 3). The agents in any group of $\mathcal{G}_{\ell-1,v}$ 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 $\mathcal{G}_{\ell-1,v}$ will become "lucky" for $a$.

(ii) Every agent in level $\ell$ observes the sub-history from $\sigma$ $(\ell-1)$-th level groups, each of which belonging to a different set $\mathcal{G}_{\ell-1,v}$. Thus, for each arm $a$, we just need one set of groups $\mathcal{G}_{\ell-1,v}$ in level $\ell - 1$ to be "lucky" for $a$ and then all agents in level $\ell$ will see sufficient arm $a$ pulls.

**Amplifying groups for boundary cases.** Recall in the three-level policy, the medium gap case (Lemma 5.6) corresponds to the case where the gap $\Delta$ is between $\Omega\left(\sqrt{1/T_1}\right)$

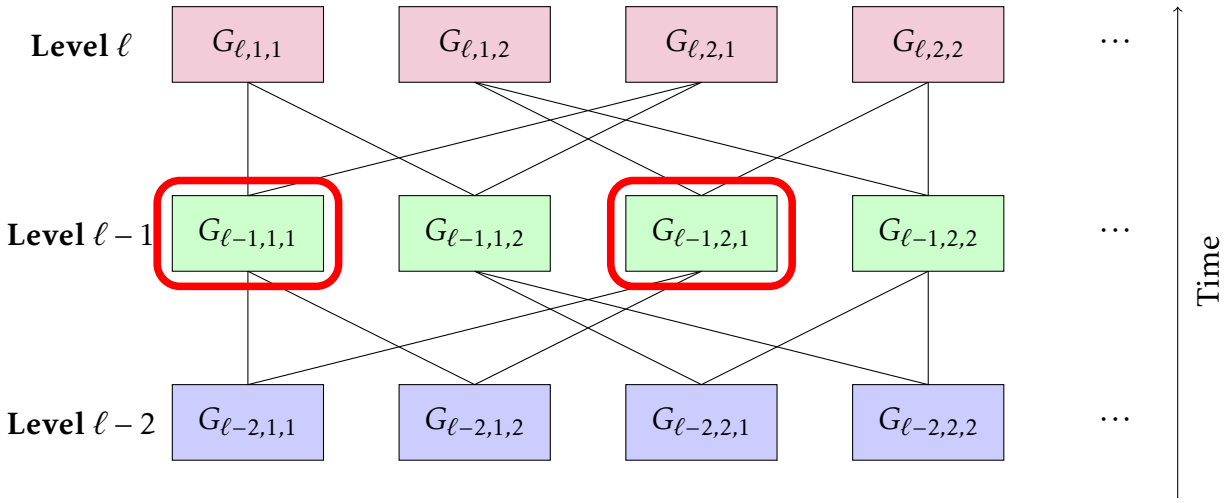


**Figure 3:** Connections between levels for the $L$-level policy, for $\sigma = 2$.

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 easily 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 $\ell \in \{2, \ldots, l-1\}$, but the issue mentioned above does not get naturally resolved since the ratios between the upper and lower bounds of $\Delta$ increase from $\Theta\left(\sqrt{\log(T)}\right)$ to $\Theta(\log(T))$, and it would require more observations from level $(\ell-2)$ to distinguish two arms at level $\ell$. 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 $G_{\ell,1,1}$, it could be that only agents in group $G_{\ell-1,1,1}$ are pulling arm $a$ and it also could be that most agents in groups $G_{\ell-1,1,1}, G_{\ell-1,1,2}, \ldots, G_{\ell-1,1,\sigma}$ 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 Γ-groups. For each $\ell \in [L], u, v \in [\sigma]$, we create a Γ-group $\Gamma_{\ell,u,v}$. Agents in $\Gamma_{\ell,u,v}$ observe the same history as the one observed by agents in $G_{\ell,u,v}$ and the number of agents in $\Gamma_{\ell,u,v}$ is $\Theta(\log(T))$ times the number of agents in $G_{\ell,u,v}$. The main difference between $G$-groups and Γ-groups is that the history of Γ-groups in level $\ell$ is not sent to agents in level $\ell+1$ but agents in higher levels. When we are in the boundary case in which we don't have good guarantees about the $(\ell+1)$-level agents' pulls, the new construction makes sure that agents in levels higher than $\ell+1$ get to see enough pulls of each arm and all pull the best arm.

**Parameters.** Aside from the global parameter $\sigma = \Theta(\log T)$ mentioned above, the structure of each level $\ell$ is determined by a parameter $T_\ell$. Specifically, we have $T_1$ full-disclosure paths in each Level-1 group. For each level $\ell = 2, \ldots, L$, each $G$-group contains $T_\ell$ agents, and each Γ-group contains $(\sigma-1)T_\ell$ agents. Parameters $T_1, \ldots, T_L$ are specified in Appendix A.5, differently for the two theorems.

# 7 Robustness

We provide several results to illustrate that our constructions are robust to small amounts of misspecification. All these results require only minor changes in the analysis, which are omitted. First, we observe that all parameters in all policies can be increased by a constant factor.[12]

**Proposition 7.1** (parameters)**.** *All results hold even if all parameters increase by at most a constant factor: specifically, parameters $(L_K^{\mathsf{FDP}}, T_1)$ for the two-level policy (Theorem 4.3), parameters $(L_K^{\mathsf{FDP}}, \sigma, T_1, T_2)$ for the three-level policy (Theorem 5.2), and parameters $(L_K^{\mathsf{FDP}}, \sigma; T_1, \ldots, T_L)$ for the L-level policy (Theorems 6.1 and 6.2).*

Let us consider a more challenging scenario when the *structure* of the communication network is altered, introducing correlation between parts of the constructions that are supposed to be isolated from one another. Recall from Example 4.8 that even a small amount of such correlation can be extremely damaging if it comes early in the game. Nevertheless, we can tolerate some undesirable correlation when it is sufficiently "local" or happens in later rounds. Informally, the existence of a local side channel between consumers does not necessarily break

[12] For the two-level policy, this is a special case of Remark 4.5. We present it here for consistency.

the regret guarantees. Families and friends can share recommendations and the reviews they've received if their social networks are sufficiently disjoint and information doesn't travel too far.

Formally, we define a generalization of the two-level policy in which the exploration level can be wired in an arbitrary way, as long as it contains sufficiently many paths that are sufficiently long and sufficiently isolated. Agents in these paths may observe some agents that lie outside of these paths, but not too many, and these outside agents may not be shared among the paths. We need a definition: for a given subset $S$ of rounds, the *span* of $S$ is the union of $S$ and all rounds $s$ that are observable in some round $t \in S$ (*i.e.*, rounds $s \leq t$ such that $s$ and $t$ are connected in the info-graph). We use quantity $L_K^{\text{FDP}}$ from Lemma 4.2.

**Proposition 7.2** (Robustness of the two-level policy)**.** *Fix some $N < T$. Consider an order-based disclosure policy such that each agent $t > N$ sees the full history: $S_t = [t-1]$. Suppose the info-graph on the first $N$ agents contains $M$ paths of length $L_K^{\text{FDP}}$ such that their spans are mutually disjoint and contain at most $2 \cdot L_K^{\text{FDP}}$ rounds each. Then*

$$\text{Reg}(T) \leq \tilde{O}_K\left(N + T/\sqrt{M}\right).$$

*In particular, we obtain* $\text{Reg}(T) \leq \tilde{O}_K\left(T^{2/3}\right)$ *when* $M = N = O(T^{2/3})$.

It is essential to bound the span size of the paths. Recall from Example 4.9 that too many "leaf agents" observed by everyone in a given full-disclosure path would rule out the natural exploration in this path.

A similar but somewhat weaker result extends to multi-level policies.

**Proposition 7.3** (Undesirable correlations in Level 1)**.** *Consider the info-graph of either multi-layer policy (from Theorem 5.2, 6.1, or 6.2). Suppose each full-disclosure path in Level 1 is replaced with subgraph $H$ which contains at most $2 \cdot L_K^{\text{FDP}}$ rounds total, includes a path of length $L_K^{\text{FDP}}$, and is connected to the rest of the info-graph via* $\max(H)$ *only. Then the corresponding theorem still holds.*

Moreover, we can handle some undesirable correlation outside of Level 1. As a proof of concept, we focus on the three-level disclosure policy, and allow each agent in Level 2 to observe some additional Level-1 agents. These agents can be chosen arbitrarily, *e.g.*, they could be the same for all Level-2 agents.

**Proposition 7.4** (undesirable correlations in Level 2)**.** *Consider the three-level policy from Theorem 5.2. Add edges to the info-graph: connect each Level-2 agent to at most $O(\sqrt{T_1})$ arbitrarily chosen agents from Level-1, where $T_1$ is the parameter from Theorem 5.2. The resulting order-based policy satisfies the guarantee in Theorem 5.2.*

# 8 Detailed comparison to prior work

Let us compare our results and modeling assumptions to those in prior work on incentivized exploration via information asymmetry. Under the strong assumptions inherent in Kremer et al. (2014) and the subsequent work, messaging policies can w.l.o.g. be reduced to multi-armed bandit algorithms which recommend an action to each agent and satisfy Bayesian incentive-compatibility (*BIC*). Hence, we will refer to this work as the *BIC incentivized exploration*.

**Trust and rationality.** We argue that order-based disclosure policies require substantially weaker trust and rationality assumptions. Several issues are in play:

*(i) Whether agents understand the announced policy.* We only need an agent to understand that she is given some unbiased history. It does not matter to the agent what subset of arrivals is covered by this subhistory, and how it is related to the other agents' subsets. This is arguably quite comprehensible, compared to a full-blown specification of a bandit algorithm.

*(ii) Whether agents trust the principal to implement the stated policy.* A third party can, at least in principle, collect subhistories from multiple agents and check them for consistency (*e.g.*, , check that arms' average rewards are within the statistical deviations), which should incentivize the principal not to manipulate the policy. Whereas bandit algorithms do not readily admit "external" sanity checks, and are extremely difficult to audit (*e.g.*, because the production code is often intertwined with many other pieces of the system, some of which may change over time or be legitimately non-public). Moreover, debugging a bandit algorithm tends to be very intricate in applications (Agarwal et al., 2017), so the implementation may deviate from the stated policy even if the principal intends otherwise. Faithfully revealing a subhistory is arguably trivial in comparison.

*(iii) Whether agents react as specified.* Agents in our model can treat the revealed subhistory as (just) a set of data-points, can exhibit a substantial amount of op-

timism or pessimism, and are not subject to the informational or cognitive load of Bayesian updates. On the other hand, agents in BIC incentivized exploration either need to trust the BIC property or verify it; the former is arguably a lot to take on faith, and the latter typically requires a sophisticated Bayesian reasoning. Moreover, agents may be irrationally averse to recommendations without any supporting information, or to the possibility of being singled out for exploration.

**Regret rates.** Like us, Mansour et al. (2020) achieve the optimal regret rate for bandit algorithms without incentives, for a constant number of actions $K$ (see Eq. (6) on page 14). Their result involves a multiplicative "constant" that can get arbitrarily large depending on the Bayesian prior. Our result similarly depends on a parameter in our choice model.

Most prior work either assumes $K = 2$ actions (*e.g.*, Kremer et al., 2014; Che and Hörner, 2018; Bimpikis et al., 2018; Bahar et al., 2016), or targets the case of constant $K$ (*e.g.*, Mansour et al., 2020, 2016). The regret bounds in prior work, as well as ours, scale exponentially in $K$. This dependence is grossly suboptimal for bandit algorithms without incentives, where one can achieve regret rates that scale as $\sqrt{K}$. A very recent, yet unpublished manuscript Sellke and Slivkins (2020) achieves BIC incentivized exploration with poly($K$) regret scaling, albeit only for independent priors and Bayesian regret (*i.e.*, regret in expectation over the Bayesian prior).

**Full disclosure and herding.** The full-disclosure policy in BIC incentivized exploration reduces to the "greedy" bandit algorithm which exploits in each round. Its herding effects are most lucidly summarized by focusing on the case of two arms. Then, if arm 1 is preferable according to the prior, the algorithm never tries arm 2 with probability at least $\mu_1^0 - \mu_2^0$, where $\mu_a^0$ is the prior mean reward of arm $a \in \{1, 2\}$. This result holds for an arbitrary priors on rewards, possibly correlated across arms. It implies very high regret (linear in $T$, the number of agents) under additional assumptions, *e.g.*, for independent priors with full support. Similar results hold for a frequentist version, where each agent chooses an arm with the highest empirical mean.[13] These results can be found in (Chapter 11.2 in Slivkins, 2019). Various weaker versions have been "folklore" for decades.

[13] For example, consider the case of two arms with Bernoulli rewards, with means $\mu_1 > \mu_2$. Assume a "warm start" such that each arm is tried $N_0$ times, $N_0 < (\mu_1 - \mu_2)^{-2}$. Then arm 2 is never chosen with probability at least an absolute constant times $\mu_1 - \mu_2$. This holds under a mild assumption on $\mu_1, \mu_2$, *e.g.*, $1/8 + \mu_1 - \mu_2 \leq \mu_2 < \mu_1 \leq 7/8$.

# 9 Conclusions

We reformulate the problem of incentivized exploration as that of designing a fixed communication network for social learning. The new model substantially mitigates trust and rationality assumptions inherent in prior work on BIC incentivized exploration (as discussed in Section 8). We achieve optimally efficient social learning, in terms of how regret rate depends on the time horizon $T$.

We start with a two-level communication network which is very intuitive and robust to misspecifications. The idea of splitting (some of) the early arrivals into many isolated "focus groups" is plausibly practical. This construction implements the explore-then-exploit paradigm from multi-armed bandits, and achieves vanishing regret. We obtain optimal regret rate via a more intricate, multi-level communication network. The conceptual challenge here is to make exploration optimally adaptive to past observation, despite the "greedy" behavior of the agents.

Incentivized exploration is rich and "multi-dimensional" problem space, in the sense that the basic model can be extended in several directions that are essentially orthogonal to each other. To wit, one could (i) consider heterogenous agents, whose idiosyncratic signals can be public or private, (ii) allow long-lived agents that strive to optimize their long-term utility, either via a suitable equilibrium concept or by running agent-side low-regret algorithms, (iii) posit some unavoidable information leakage, *e.g.*, according to a pre-specified social network, and (iv) optimize the dependence on the number of arms and the agents' beliefs. All these directions are extremely interesting, and some of them have been studied, yet they are not well-understood even under the strong assumptions of BIC incentivized exploration.

# Acknowledgments

We would like to thank Robert Kleinberg for discussions in the early stage of this project, Ian Ball for thoughtful comments on a draft; seminar attendees at Columbia, Michigan, NYU, Princeton, Yale, the Simons Institute for the Theory of Computing, Stanford; and workshop attendees at the Arizona State University Economic Theory Conference, the Stony Brook Game Theory Center, and the UBC-HKU Summer Theory Conference.

# Bibliography


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

Daron Acemoglu, Kostas Bimpikis, and Asuman Ozdaglar. Dynamics of information exchange in endogenous social networks. *Theoretical Economics*, 9:41–97, 2014.

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Learning From Reviews: The Selection Effect and the Speed of Learning, 2019. Working paper. Revise and resubmit in *Econometrica*.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt, 2017. Techical report at `arxiv.org/abs/1606.03966`.

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. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. on 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 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):13:1–13:55, 2018. Preliminary version in *FOCS 2013*.

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

Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Social learning and the innkeeper's challenge. In *ACM Conf. on Economics and Computation (ACM-EC)*, pages 153–170, 2019.

Abhijit Banerjee and Drew Fudenberg. Word-of-mouth learning. *Games and Economic Behavior*, 46:1–22, 2004.

Abhijit V. Banerjee. A simple model of herd behavior. *Quarterly Journal of Economics*, 107: 797–817, 1992.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 2020. Working paper available on arxiv.org since 2017.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, March 2019.

Dirk Bergemann and Juuso Välimäki. Bandit Problems. In Steven Durlauf and Larry Blume, editors, *The New Palgrave Dictionary of Economics, 2nd ed.* Macmillan Press, 2006.

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(6):1407–1420, 2009.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100 (5):992–1026, 1992.

Kostas Bimpikis, Yiangos Papanastasiou, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64(4):1477–1973, 2018.

Celen Bogachan and Shachar Kariv. Observational learning under imperfect information. *Games and Economic Behavior*, 47:72–86, 2004.

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 Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122, 2012. Published with *Now Publishers* (Boston, MA, USA). Also available at https://arxiv.org/abs/1204.5721.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, UK, 2006.

Arun G Chandrasekhar, Horacio Larreguy, and Juan Pablo Xandri. Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32, 2020.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *Quarterly Journal of Economics*, 133(2):871–925, 2018. Working paper since 2013, titled 'Optimal design for social learning'.

Bangrui Chen, Peter I. Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In *Conf. on Learning Theory (COLT)*, pages 798–818, 2018.

Lee Cohen and Yishay Mansour. Optimal algorithm for bayesian incentive-compatible exploration. In *ACM Conf. on Economics and Computation (ACM-EC)*, pages 135–151, 2019.

Krishna Dasaratha and Kevin He. An experiment on network density and sequential learning. *arXiv preprint arXiv:1909.02220*, 2019.

Krishna Dasaratha and Kevin He. Aggregative efficiency of bayesian learning in networks. *working paper*, 2020.

Morris H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (ACM-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.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011. The first edition, single-authored by John Gittins, has been published in 1989.

Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49, 2010.

Benjamin Golub and Evan D. Sadler. Learning in social networks. In Yann Bramoullé, Andrea Galeotti, and Brian Rogers, editors, *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.

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*.

Johannes Hörner and Andrzej Skrzypacz. Learning, experimentation, and information design. In Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, editors, *Advances in Economics and Econometrics: Eleventh World Congress*, volume 1 of *Econometric Society Monographs*, page 63–98. Cambridge University Press, 2017.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Steven Wu. Bayesian exploration with heterogenous agents. In *The Web Conference (formerly known as* WWW*)*, pages 751–761, 2019.

Maxim Ivanov. Optimal signals in bayesian persuasion mechanisms. *working paper*, 2015.

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

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

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)*, pages 594–605, 2003.

Robert D. Kleinberg, Bo Waggoner, and E. Glen Weyl. Descending price optimally coordinates search. In *17th ACM Conf. on Economics and Computation (ACM-EC)*, pages 23–24, 2016.

Anton Kolotilin. Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635, 2018.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *J. of Political Economy*, 122(5):988–1012, 2014. Preliminary version in *ACM EC 2013*.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020. Versions available at `https://banditalgs.com/` since 2018.

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.

Ilan Lobel and Evan Sadler. Information diffusion in networks through social learning. *Theoretical Economics*, 10(3):807–851, 2015.

Mark J Machina and Marciano Siniscalchi. Ambiguity and ambiguity aversion. In *Handbook of the Economics of Risk and Uncertainty*, volume 1, pages 729–807. Elsevier, 2014.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Steven Wu. Bayesian exploration: Incentivizing exploration in Bayesian games, 2016. Working paper (2016-2018). Avalable at https://arxiv.org/abs/1602.07570. Prelimiary version in *ACM EC 2016*. Revise and resubmit in *Operations Research*.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020. Preliminary version in *ACM EC 2015*.

Jeffrey Mensch. Monotone persuasion. *Available at SSRN 3265980*, 2019.

Afshin Nikzad. Persuading a pessimist: Simplicity and robustness. *working paper*, 2019.

Sanket Patil and Yuval Salant. Persuading statisticians. *working paper*, 2020.

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. Greedy algorithm almost dominates in smoothed contextual bandits, 2018. Working paper. Preliminary version in *COLT 2018*.

Mark Sellke and Aleksandrs Slivkins. Sample complexity of insentivized exploration, 2020. Working paper, available at https://arxiv.org/abs/2002.00558.

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.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at https://arxiv.org/abs/1904.07272.

Lones Smith and Peter Sørensen. Pathological outcomes of observational learning. *Econometrica*, 68:371–398, 2000.

Ivo Welch. Sequential sales, learning, and cascades. *The Journal of finance*, 47:695–732, 1992.

# Appendix A Proofs

## A.1 Preliminaries

We use the standard concentration and anti-concentration inequalities: respectively, Chernoff Bounds and Berry-Esseen Theorem. The former states that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the average of $n$ independent random variables $X_1, \ldots, X_n$, converges to its expectation quickly. The latter states that the CDF of an appropriately scaled average $\bar{X}$ converges to the CDF of the standard normal distribution pointwise. In particular, the average strays far enough from its expectation with some guaranteed probability. The theorem statements are as follows:

**Theorem A.1.** *Fix $n$. Let $X_1, \ldots, X_n$ be independent random variables, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:*

*(a)* (Chernoff Bounds) *Assume $X_i \in [0,1]$ for all $i$. Then*

$$\Pr[|\bar{X} - \mathbb{E}[\bar{X}]| > \varepsilon] \leq 2\exp(-2n\varepsilon^2).$$

*(b)* (Berry-Esseen Theorem) *Assume $X_1, \ldots, X_n$ are identically distributed, with*

$$\sigma^2 := \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] \quad \text{and} \quad \rho := \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^3] < \infty.$$

*Let $F_n$ be the cumulative distribution function of $\frac{(\bar{X} - \mathbb{E}[\bar{X}])\sqrt{n}}{\sigma}$ and $\Phi$ be the cumulative distribution function of the standard normal distribution.*

$$|F_n(x) - \Phi(x)| \leq \frac{\rho}{2\sigma^3\sqrt{n}} \quad \forall x \in \mathbb{R}.$$

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

## A.2 The two-level policy: proof of Theorem 4.3

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 $T_1 \geq \frac{4(L_K^{\mathsf{FDP}})^2}{(p_K^{\mathsf{FDP}})^2}\log(T)$. 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 $T_1 \cdot L_K^{\mathsf{FDP}}$ rounds will get total regret at most $T_1 \cdot L_K^{\mathsf{FDP}}$. We focus on bounding the regret from the second level of $T - T_1 \cdot L_K^{\mathsf{FDP}}$ 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 $N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$. As long as we set $T_1 \geq \frac{4(L_K^{\mathsf{FDP}})^2}{(p_K^{\mathsf{FDP}})^2} \log(T)$, this implies that the number of arm $a$ pulls is then at least $N_{K,a}^{\mathsf{FDP}} T_1/2$. Define $W_1$ to be the intersection of all these events (i.e. $W_1 = \bigcap_a W_1^a$). By Lemma 4.7, 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 $\mathcal{D}_a$. 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 $\sqrt{\frac{2\log(T)}{\tau}}$ 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

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

By $W_1$ and Assumption 3.1, we have

$$|\bar{\mu}_a^t - \hat{\mu}_a^t| \leq \frac{C_{\mathtt{est}}}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1/2}}.$$

Therefore,

$$|\hat{\mu}_a^t - \mu_a| \leq \sqrt{\frac{2\log(T)}{N_{K,a}^{\mathsf{FDP}} T_1/2}} + \frac{C_{\mathtt{est}}}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1/2}} \leq 3\sqrt{\frac{\log(T)}{p_K^{\mathsf{FDP}} T_1}}.$$

So the second-level agents will pick an arm $a$ which has $\mu_a$ at most $6\sqrt{\frac{\log(T)}{p_K^{\mathsf{FDP}} T_1}}$ away from $\mu_1$. To sum up, the total regret is at most

$$T_1 \cdot L_K^{\mathsf{FDP}} + T \cdot (1 - \Pr[W_1 \cap W_2]) + T \cdot 6\sqrt{\frac{\log(T)}{p_K^{\mathsf{FDP}} T_1}}.$$

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

## A.3 The "global" counterexample: proof for Example 4.8

We consider three events, denoted $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$. Event $\mathcal{E}_1$ is that after the first $N_1 = 2$ rounds, arm 1 has empirical mean at most $\mu' < \mu_2$ and arm 2 empirical mean at least $\mu_2$. (The proof can work for other constant $N_1$, too.) We pick $\mu'$ such that $\mu_2 - \mu' = \Omega(1)$. Event $\mathcal{E}_2$ focuses on the next $N - N_1$ rounds. It asserts that arm 2 is the only one chosen in these rounds, and the empirical mean in any prefix of these rounds is at least $\mu_2$. Event $\mathcal{E}_3$ is that the last $T - N$ agents all choose arm 2.

We lower-bound $\Pr[\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3]$ by a positive constant by considering $\Pr[\mathcal{E}_1]$, $\Pr[\mathcal{E}_2 \mid \mathcal{E}_1]$ and $\Pr[\mathcal{E}_3 \mid \mathcal{E}_1, \mathcal{E}_2]$. First, $\mathcal{E}_1$ happens with a constant probability as arm 1 getting 0 in its first pull and arm 2 getting 1 in its first pull is a sub case of $\mathcal{E}_1$.

Now we condition on $\mathcal{E}_1$ happening. We show that $\mathcal{E}_2$ happens with a positive-constant probability. We focus on the case when the first $N_2$ pulls of arm 2 in rounds $\{N_1+1, \ldots, N\}$ are all 1's for some large enough constant $N_2$ and then use Chernoff bound and union bound on the rest $N - N_1 - N_2$ pulls.

Now we condition on $\mathcal{E}_1$ and $\mathcal{E}_2$. We consider a "reward tape" generating rewards of arm 2, where the $t$-th "cell" in the tape corresponds to the reward of arm 2 in round $t$ if this arm is chosen in this round. For each $t > N$, let $C_t$ be the subset of cells in the tape that correspond to rounds $S_t \cap (N, T]$, where $S_t$ is the set of rounds observable by agent $t$. We can show that with very high probability, the empirical mean over $C_t$ is larger than $\mu'$ for all $t$. Let us focus on this event, call it $\mathcal{E}_{\text{tape}}$. We show that under $\mathcal{E}_{\text{tape}}$, each agents $t > N$ chooses arm 2, using induction on $t$. This is because $C_t$, together with the history of the first $N$ rounds, is exactly the subhistory seen by agent $t$, if all agents in round $\{N+1, \ldots, t-1\}$ pull arm 2.

## A.4 The three-level policy: proof of Theorem 5.2

### High-probability events

The following lemmas can be derived from combining Lemma 4.7 and union bound.

**Lemma A.2** (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*

$$\left[N^{\text{FDP}}_{K,a} T_1 - L^{\text{FDP}}_K \sqrt{T_1 \log(T)}, N^{\text{FDP}}_{K,a} T_1 + L^{\text{FDP}}_K \sqrt{T_1 \log(T)}\right],$$

*where $N^{\text{FDP}}_{K,a}$ is the expected number of arm $a$ pulls in a $full-disclosure$ path run of length $L^{\text{FDP}}_K$. Then $\Pr[W_1] \geq 1 - \frac{4\sigma}{T^2}$.*

*Proof of Lemma A.2.* 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 $N^{\text{FDP}}_{K,a} T_1 - L^{\text{FDP}}_K \sqrt{T_1 \log(T)}$ and $N^{\text{FDP}}_{K,a} T_1 + L^{\text{FDP}}_K \sqrt{T_1 \log(T)}$. By Lemma 4.7

$$\Pr[W_1^{a,s}] \geq 1 - 2\exp(-2\log(T)) \geq 1 - 2/T^2.$$

By union bound, the intersection of all these events, $\bigcap_{a,s} W_1^{a,s}$, has probability at least $1 - \frac{4\sigma}{T^2}$. □

To state the events, it will be useful to think of a hypothetical reward tape $\mathcal{T}^1_{s,a}$ of length $T$ for each group $s$ and arm $a$, with each cell independently sampled from $\mathcal{D}_a$. 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 A.3** (Concentration of empirical means in the first level)**.** *For any $\tau_1, \tau_2 \in [T]$ such that $\tau_1 < \tau_2$, $s \in [\sigma]$, 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 $\mathcal{T}^1_{a,s}$ is at most $\sqrt{\frac{2\log(T)}{\tau_2-\tau_1+1}}$ 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 of Lemma A.3.* By Chernoff bound,

$$\Pr[W_2^{s,a,\tau_1,\tau_2}] \geq 1 - 2\exp(-4\log(T)) \geq 1 - 2/T^4.$$

By union bound, we have $\Pr[W_2] \geq 1 - 4\sigma/T^2$. □

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 A.4** (Co-occurence of high and low deviations in this first level)**.** *For any group $s \in [\sigma]$, any arm $a$, let $\tilde{\mu}_{a,s}$ be the empirical mean of the first $N^{\mathsf{FDP}}_{K,a} T_1$ cells in tape $\mathcal{T}^1_{a,s}$. Let $W_3^{s,a,high}$ be the event $\tilde{\mu}_{a,s} \geq \mu_a + 1/\sqrt{N^{\mathsf{FDP}}_{K,a} T_1}$ and let $W_3^{s,a,low}$ be the event that $\tilde{\mu}_{a,s} \leq \mu_a - 1/\sqrt{N^{\mathsf{FDP}}_{K,a} T_1}$. Let $W_3$ be the event that for every $a \in \{1,2\}$, there exists a group $s_a \in [\sigma]$ in the first level such that both $W_3^{s_a,a,high}$ and $W_3^{s_a,3-a,low}$ occur. Then $\Pr[W_3] \geq 1 - 2/T$.*

*Proof of Lemma A.4.* By Berry-Esseen Theorem and $\mu_a \in [1/3, 2/3]$, we have for any $a$,

$$\Pr[W_3^{s,a,high}] \geq (1 - \Phi(1/2)) - \frac{5}{\sqrt{N^{\mathsf{FDP}}_{K,a} T_1}} > 1/4.$$

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

$$\Pr[W_3^{s,a,low}] > 1/4.$$

Since $W_3^{s,a,high}$ is independent with $W_3^{s,3-a,low}$, we have

$$\Pr[W_3^{s,a,high} \cap W_3^{s,3-a,low}] = \Pr[W_3^{s,a,high}] \cdot \Pr[W_3^{s,3-a,low}] > (1/4)^2 = 1/16.$$

Notice that $(W_3^{s,a,high} \cap W_3^{s,3-a,low})$ 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 A.5** (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 of Lemma A.5.* For any arm $a$, let's imagine a hypothetical tape of length $T$, with each cell independently sampled from $\mathcal{D}_a$. 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 $W = \bigcap_{i=1}^4 W_i$ 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 $W$.

### 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 A.6.** *For the s-th first-level group and arm a, define* $\bar{\mu}_a^{1,s}$ *to be the empirical mean of arm a pulls in this group. If* $W$ *holds, then*

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

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

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

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

**Lemma A.7.** *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*

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

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

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

*Proof.* The events $W_1$ and $W_4^{a,\tau}$ for $\tau \ge (N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)})\sigma$ together imply that

$$|\bar{\mu}_a - \mu_a| \le \sqrt{\frac{2\log(T)}{\sigma\left(N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)}\right)}} \le \sqrt{\frac{4\log(T)}{\sigma N_{K,a}^{\mathsf{FDP}} T_1}}.$$

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

**Lemma A.8.** *For the $s$-th first-level group and arm $a$, define $\bar{\mu}_a^{1,s}$ 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*

$$\bar{\mu}_a^{1,s_a} > \mu_a + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}} \quad \textit{and,} \quad \bar{\mu}_{3-a}^{1,s_a} < \mu_{3-a} - \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1}}.$$

*Proof.* For each $a \in \{1,2\}$, $W_3$ implies that there exists $s_a$ such that both $W_3^{s_a,a,high}$ and $W_3^{s_a,3-a,low}$ happen. The events $W_3^{s_a,a,high}$, $W_1$, $W_2^{s_a,a,\tau,N_{K,a}^{\mathsf{FDP}} T_1}$ for $\tau = N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)} + 1, \ldots, N_{K,a}^{\mathsf{FDP}} T_1 - 1$ and $W_2^{s_a,a,N_{K,a}^{\mathsf{FDP}} T_1,\tau}$ for $\tau = N_{K,a}^{\mathsf{FDP}} T_1, \ldots, N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)}$ together imply that

$$\begin{aligned}\bar{\mu}_a^{1,s_a} &\ge \mu_a + \left(N_{K,a}^{\mathsf{FDP}} T_1 \cdot \frac{1}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}} - L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)} \cdot \sqrt{\frac{2\log(T)}{L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)}}}\right) \cdot \frac{1}{N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}}\sqrt{T_1\log(T)}} \\ &> \mu_a + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}}.\end{aligned}$$

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

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

□

Now we proceed to the case analysis.

*Proof of Lemma 5.5 (Large gap case).* Observe that for any group $s$ in the first level, the em-

pirical means satisfy

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

For any agent $t$ in the $s$-th second-level group, by Assumption 3.1, we have

$$\begin{aligned}\hat{\mu}_1^t - \hat{\mu}_2^t &> \bar{\mu}_1^{1,s} - \bar{\mu}_2^{1,s} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,2}^{\mathsf{FDP}} T_1/2}} \\ &\geq \sqrt{\frac{4\log(T)}{N_{K,1}^{\mathsf{FDP}} T_1}} + \sqrt{\frac{4\log(T)}{N_{K,2}^{\mathsf{FDP}} T_1}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,2}^{\mathsf{FDP}} T_1/2}} > 0.\end{aligned}$$

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 $\bar{\mu}_a$ is the empirical mean of arm $a$ in the history they see. We have

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

Similarly as above, by Assumption 3.1, we know $\hat{\mu}_1^t - \hat{\mu}_2^t > 0$ for any agent $t$ in the third level. Therefore, the agents in the third-level group will all pull arm 1. □

*Proof of Lemma 5.6 (Medium gap case).* Recall $\bar{\mu}_a$ is the empirical mean of arm $a$ in the first two levels. We have

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

For any agent $t$ in the third level, by Assumption 3.1, we have

$$\begin{aligned}\hat{\mu}_1^t - \hat{\mu}_2^t &> \bar{\mu}_1 - \bar{\mu}_2 - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} \\ &\geq \sqrt{\frac{4\log(T)}{\sigma N_{K,1}^{\mathsf{FDP}} T_1}} + \sqrt{\frac{4\log(T)}{\sigma N_{K,2}^{\mathsf{FDP}} T_1}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} \\ &> 0.\end{aligned}$$

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

*Proof of Lemma 5.7 (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 A.8. We have

$$\begin{aligned}\bar{\mu}_a^{1,s_a} - \bar{\mu}_{3-a}^{1,s_a} &> \mu_a + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}} - \mu_{3-a} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1}} \\ &> \frac{1}{4\sqrt{N_{K,1}^{\mathsf{FDP}} T_1}} + \frac{1}{4\sqrt{N_{K,2}^{\mathsf{FDP}} T_1}} - 2\left(\sqrt{\frac{4\log(T)}{\sigma N_{K,1}^{\mathsf{FDP}} T_1}} + \sqrt{\frac{4\log(T)}{\sigma N_{K,2}^{\mathsf{FDP}} T_1}}\right) \\ &\geq \frac{1}{8\sqrt{N_{K,1}^{\mathsf{FDP}} T_1}} + \frac{1}{8\sqrt{N_{K,2}^{\mathsf{FDP}} T_1}}.\end{aligned}$$

For any agent $t$ in the $s_a$-th second-level group, by Assumption 3.1, we have

$$\begin{aligned}\hat{\mu}_a^t - \hat{\mu}_{3-a}^t &> \bar{\mu}_a^{1,s_a} - \bar{\mu}_{3-a}^{1,s_a} - \frac{C_{\mathsf{est}}}{\sqrt{N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathsf{est}}}{\sqrt{N_{K,2}^{\mathsf{FDP}} T_1/2}} \\ &\geq \frac{1}{8\sqrt{N_{K,1}^{\mathsf{FDP}} T_1}} + \frac{1}{8\sqrt{N_{K,2}^{\mathsf{FDP}} T_1}} - \frac{C_{\mathsf{est}}}{\sqrt{N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathsf{est}}}{\sqrt{N_{K,2}^{\mathsf{FDP}} T_1/2}} \\ &> 0.\end{aligned}$$

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 3.1, 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. □

## A.5 The multi-level policy

In this subsection, we analyze our $L$-level policy for $L > 3$, proving Theorems 6.1 and 6.2. We first analyze it for the case of $K = 2$ arms. The bulk of the analysis, joint for both theorems, is presented in Appendix A.5.1. We provide two different endings where the details differ: Appendix A.5.2 and Appendix A.5.3, respectively. We extend the analysis to $K > 2$ arms in Appendix A.5.4.

The parameters are set as follows. In Theorem 6.1, recall from the theorem statement that we restrict $L$ to be at most $L_{\max} = \Theta(\log\log T)$. Specifically, we define

$$L_{\max} = \log\left(\frac{\ln T}{\log \sigma^4}\right).$$

The group number parameter $\sigma$ is set at $\sigma = 2^{10}\log(T)$ for both theorems. Parameters $T_1, \ldots, T_L$ are specified differently for the two theorems, see (8), (9) and (10).

Let us recap the construction of the $L$-level policy. There are two types of groups: $G$-groups and $\Gamma$-groups. Each level has $\sigma^2$ $G$-groups. Label the $G$-groups in the $\ell$-th level as $G_{\ell,u,v}$ for $u,v \in [\sigma]$. Level 2 to level $L$ also have $\sigma^2$ $\Gamma$-groups. Label the $\Gamma$-groups in the $\ell$-th level as $\Gamma_{\ell,u,v}$ 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 $L_K^{\mathsf{FDP}}$ rounds in parallel. For $\ell \geq 2$, there are $T_\ell$ agents in group $G_{\ell,u,v}$ and there are $T_\ell(\sigma-1)$ agents in group $\Gamma_{\ell,u,v}$.

The info-graph is defined as follows. Agents in the first level only observe the history defined in the full-disclosure path run. For agents in group $G_{\ell,u,v}$ with $\ell \geq 2$, they observe all the history in the first $\ell-2$ levels (both $G$-groups and $\Gamma$-groups) and history in group $G_{\ell-1,v,w}$ for all $w \in [\sigma]$. Agents in group $\Gamma_{\ell,u,v}$ observe the same history as agents in group $G_{\ell,u,v}$.

### A.5.1 Joint analysis for $K = 2$ arms

The bulk of the analysis is joint for Theorems 6.1 and 6.2. While the parameters $T_\ell$ are set differently for the two theorems, we will only assume

$$T_1 \leq \sigma^4 \leq \frac{T_\ell}{T_{\ell-1}} \quad \text{for} \quad \ell \in \{2, \dots, L-1\}, \tag{7}$$

which will hold for both parameter settings. Wlog we assume $\mu_1 \geq \mu_2$ as the recommendation policy is symmetric to both arms.

Similarly as the proof of Theorem 5.2, we start with some "clean events".

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

  For $a \in \{1,2\}$, define $N_{K,a}^{\mathsf{FDP}}$ to be the expected number of arm $a$ pulls in one run of full-disclosure path used in the first level. By Lemma 4.2, we know $p_K^{\mathsf{FDP}} \leq N_{K,a}^{\mathsf{FDP}} \leq L_K^{\mathsf{FDP}}$ For group $G_{1,u,v}$, define $W_1^{a,u,v}$ to be the event that the number of arm $a$ pulls in this group is between $N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$ and $N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$. By Chernoff bound,

  $$\Pr[W_1^{a,u,v}] \geq 1 - 2\exp(-2\log(T)) \geq 1 - 2/T^2.$$

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

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

- **Concentration of the empirical mean for arm $a$ 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 $\ell$-th level for $\ell \geq 2$:**
  For $2 \leq \ell \leq L-1$, $u \in [\sigma]$ and each arm $a$, define $n^{\ell,u,a}$ to be the number of arm $a$ pulls in groups $G_{\ell,u,1},...,G_{\ell,u,\sigma}$. Define $W_3^{\ell,u,a,high}$ as the event that $n^{\ell,u,a} \geq T_\ell$ implies the empirical mean of arm $a$ pulls in group $G_{\ell,u,1},...,G_{\ell,u,\sigma}$ is at least $\mu_a + 1/\sqrt{n^{\ell,u,a}}$. Define $W_3^{\ell,u,a,low}$ as the event that $n^{\ell,u,a} \geq T_\ell$ implies the empirical mean of arm $a$ pulls in group $G_{\ell,u,1},...,G_{\ell,u,\sigma}$ is at most $\mu_a - 1/\sqrt{n^{\ell,u,a}}$.

  Define $H_\ell$ to be random variable the history of all agents in the first $\ell-1$ levels and which agents are chosen in the $\ell$-th level. Let $h_\ell$ be some realization of $H_\ell$. Notice that once we fix $H_\ell$, $n^{\ell,u,a}$ is also fixed.

  Now consider $h_\ell$ to be any possible realized value of $H_\ell$. If fixing $H_\ell = h_\ell$ makes $n^{\ell,u,a} < T_\ell$, then $\Pr[W_3^{\ell,u,a,high} | H_\ell = h_\ell] = 1$ If fixing $H_\ell = h_\ell$ makes $n^{\ell,u,a} \geq T_\ell$, by Berry-Esseen Theorem and $\mu_a \in [1/3, 2/3]$, we have

$$\Pr[W_3^{\ell,u,a,high} | H_\ell = h_\ell] \geq (1 - \Phi(1/2)) - \frac{5}{\sqrt{T_\ell}} > 1/4.$$

  Similarly we also have

$$\Pr[W_3^{\ell,u,a,low} | H_\ell = h_\ell] > 1/4$$

  Since $W_3^{\ell,u,a,high}$ is independent with $W_3^{\ell,u,3-a,low}$ when fixing $H_\ell$, we have

$$\Pr[W_3^{\ell,u,a,high} \cap W_3^{\ell,u,3-a,low} | H_\ell = h_\ell] > (1/4)^2 = 1/16.$$

  Now define $W_3^{\ell,a} = \bigcup_u (W_3^{\ell,u,a,high} \cap W_3^{\ell,u,3-a,low})$. Since $(W_3^{\ell,u,a,high} \cap W_3^{\ell,u,3-a,low})$ are independent across different $u$'s when fixing $H_\ell = h_\ell$, we have

$$\Pr[W_3^{\ell,a} | H_\ell = h_\ell] \geq 1 - (1 - 1/16)^\sigma \geq 1 - 1/T^2.$$

  Since this holds for all $h_\ell$'s, we have $\Pr[W_3^{\ell,a}] \geq 1 - 1/T^2$. Finally define $W_3 = \bigcap_{\ell,a} W_3^{\ell,a}$. 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 $G_{1,u,1},...,G_{1,u,\sigma}$ 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 $N_{K,a}^{\mathsf{FDP}} T_1 \sigma$ pulls of arm $a$ in the tape has empirical mean at least $\mu_a + 1/\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}$ and define $W_4^{u,a,low}$ to be the event that first $N_{K,a}^{\mathsf{FDP}} T_1 \sigma$ pulls of arm $a$

in the tape has empirical mean at most $\mu_a - 1/\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}$. By Berry-Esseen Theorem and $\mu_a \in [1/3, 2/3]$, we have

$$\Pr[W_4^{u,a,high}] \geq (1 - \Phi(1/2)) - \frac{5}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} > 1/4.$$

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 $\bigcup_u (W_4^{u,a,high} \cap W_4^{u,3-a,low})$. Notice that $(W_4^{u,a,high} \cap W_4^{u,3-a,low})$ are independent across different $u$'s. So we have

$$\Pr[W_4^a] \geq 1 - (1 - 1/16)^\sigma \geq 1 - 1/T^2.$$

Finally we define $W_4$ as $\bigcap_a W_4^a$. 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 $N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$ 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 $\epsilon_0 = 1$, $\epsilon_1 = \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}}$ and $\epsilon_\ell = 1/(4\sqrt{T_\ell \sigma})$ for $\ell \geq 2$.

**Claim A.9.** *For any arm $a$ and $2 \leq \ell \leq L$, if $\mu_1 - \mu_2 \leq \epsilon_{\ell-1}$, then for any $u \in [\sigma]$, there are at least $T_\ell$ pulls of arm $a$ in groups $G_{\ell,u,1}, G_{\ell,u,2}, ..., G_{\ell,u,\sigma}$ and there are at least $T_\ell \sigma(\sigma - 1)$ pulls of arm $a$ in the $\ell$-th level $\Gamma$-groups.*

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

We prove the above via induction on $\ell$. We start by the base case when $\ell = 2$. For each arm $a$, $W_4$ implies there exists $u_a$ such that $W_4^{u_a,a,high}$ and $W_4^{u_a,3-a,low}$ happen. For an agent $t$ in groups $G_{2,1,u_a}, ..., G_{2,\sigma,u_a}$ and $\Gamma_{2,1,u_a}, ..., \Gamma_{2,\sigma,u_a}$. $W_4^{u_a,a,high}$, $W_1^{a,u_a,v}$ and $W_2$ together imply

that

$$\begin{aligned}\bar{\mu}_a^t &\geq \mu_a + \left( N_{K,a}^{\mathsf{FDP}} T_1 \sigma \cdot \frac{1}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)} \sigma \cdot \sqrt{\frac{3\log(T)}{L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)} \sigma}} \right) \\ &\quad \cdot \frac{1}{(N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}) \sigma} \\ &> \mu_a + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}}.\end{aligned}$$

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

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

Then we have

$$\begin{aligned}\bar{\mu}_a^t - \bar{\mu}_{3-a}^t &> \mu_a - \mu_{3-a} + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}} \\ &\geq -\epsilon_1 + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}} \\ &\geq \frac{1}{8\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{8\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}}.\end{aligned}$$

By Assumption 3.1, we have

$$\begin{aligned}\hat{\mu}_a^t - \hat{\mu}_{3-a}^t &> \bar{\mu}_a^t - \bar{\mu}_{3-a}^t - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma/2}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma/2}} \\ &> \frac{1}{8\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{8\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma/2}} - \frac{C_{\mathrm{est}}}{\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma/2}} \\ &> 0.\end{aligned}$$

The last inequality holds since $C_{\mathrm{est}}$ is a small enough constant defined in Assumption 3.1. Therefore we know agents in groups $G_{2,1,u_a}, ..., G_{2,\sigma,u_a}$ and $\Gamma_{2,1,u_a}, ..., \Gamma_{2,\sigma,u_a}$ all pull arm $a$.

Now we consider the case when $\ell > 2$ and assume the claim is true for smaller $\ell$'s. For each arm $a$, $W_3$ implies that there exists $u_a$ such that $W_3^{\ell-1,u_a,a,high}$ and $W_3^{\ell-1,u_a,3-a,low}$ happen. Recall $n^{\ell-1,u_a,a}$ is the number of arm $a$ pulls in groups $G_{\ell-1,u_a,1}, ..., G_{\ell-1,u_a,\sigma}$. The induction hypothesis implies that $n^{\ell-1,u_a,a} \geq T_{\ell-1}$. $W_3^{\ell-1,u_a,a,high}$ together with $n^{\ell-1,u_a,a} \geq T_{\ell-1}$ implies that the empirical mean of arm $a$ pulls in group $G_{\ell-1,u_a,1}, ..., G_{\ell-1,u_a,\sigma}$ is at least $\mu_a + 1/\sqrt{n^{\ell-1,u_a,a}}$. For any agent $t$ in groups $G_{\ell,1,u_a}, ..., G_{\ell,\sigma,u_a}$ and $\Gamma_{\ell,1,u_a}, ..., \Gamma_{\ell,\sigma,u_a}$, it observes

history of groups $G_{\ell-1,u_a,1},...,G_{\ell-1,u_a,\sigma}$ and all groups in levels below level $\ell-1$. Notice that the groups in the first $\ell-2$ levels have at most $(T_1 L_K^{\mathsf{FDP}} + T_2 + \cdots + T_{\ell-2})\sigma^3 \le T_{\ell-1}/(12\log(T)) \le n^{\ell-1,u_a,a}/(12\log(T))$ agents. By $W_2$, we have

$$\begin{aligned}\bar{\mu}_a^t &\ge \mu_a + \left( n^{\ell-1,u_a,a} \cdot \frac{1}{\sqrt{n^{\ell-1,u_a,a}}} - (T_1 L_K^{\mathsf{FDP}} + T_2 + \cdots + T_{\ell-2})\sigma^3 \cdot \sqrt{\frac{3\log(T)}{(T_1 L_K^{\mathsf{FDP}} + T_2 + \cdots + T_{\ell-2})\sigma^3}} \right) \\ &\quad \cdot \frac{1}{n^{\ell-1,u_a,a} + (T_1 L_K^{\mathsf{FDP}} + T_2 + \cdots + T_{\ell-2})\sigma^3} \\ &> \mu_a + \frac{1}{4\sqrt{n^{\ell-1,u_a,a}}}.\end{aligned}$$

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

$$\bar{\mu}_{3-a}^t < \mu_{3-a} - \frac{1}{4\sqrt{n^{\ell-1,u_a,3-a}}}.$$

Then we have

$$\begin{aligned}\bar{\mu}_a^t - \bar{\mu}_{3-a}^t &> \mu_a - \mu_{3-a} + \frac{1}{4\sqrt{n^{\ell-1,u_a,a}}} + \frac{1}{4\sqrt{n^{\ell-1,u_a,3-a}}} \\ &\ge -\epsilon_{\ell-1} + \frac{1}{4\sqrt{n^{\ell-1,u_a,a}}} + \frac{1}{4\sqrt{n^{\ell-1,u_a,3-a}}} \\ &\ge \frac{1}{8\sqrt{n^{\ell-1,u_a,a}}} + \frac{1}{8\sqrt{n^{\ell-1,u_a,3-a}}}.\end{aligned}$$

The last inequality holds because $n^{\ell-1,u_a,a}$ and $n^{\ell-1,u_a,3-a}$ are at most $T_{\ell-1}\sigma$. By Assumption 3.1, we have

$$\begin{aligned}\hat{\mu}_a^t - \hat{\mu}_{3-a}^t &> \bar{\mu}_a^t - \bar{\mu}_{3-a}^t - \frac{C_{\mathrm{est}}}{\sqrt{n^{\ell-1,u_a,a}}} - \frac{C_{\mathrm{est}}}{\sqrt{n^{\ell-1,u_a,3-a}}} \\ &> \frac{1}{8\sqrt{n^{\ell-1,u_a,a}}} + \frac{1}{8\sqrt{n^{\ell-1,u_a,3-a}}} - \frac{C_{\mathrm{est}}}{\sqrt{n^{\ell-1,u_a,a}}} - \frac{C_{\mathrm{est}}}{\sqrt{n^{\ell-1,u_a,3-a}}} \\ &> 0.\end{aligned}$$

The last inequality holds since $C_{\mathrm{est}}$ is a small enough constant defined in Assumption 3.1. Therefore agents in groups $G_{\ell,1,u_a},...,G_{\ell,\sigma,u_a}$ and $\Gamma_{\ell,1,u_a},...,\Gamma_{\ell,\sigma,u_a}$ all pull arm $a$. □

**Claim A.10.** *For any $2 \le \ell \le L$, if $\epsilon_{\ell-1}\sigma \le \mu_1 - \mu_2 < \epsilon_{\ell-2}\sigma$, there are no pulls of arm 2 in groups with level $\ell,...,L$.*

*Proof.* We argue in 2 cases $\epsilon_{\ell-1}\sqrt{\sigma} \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}$ for $\ell \ge 2$ and $\epsilon_{\ell-2} \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}\sqrt{\sigma}$ for $\ell > 2$. Since our recommendation policy's first level is slightly different from other levels, we need to argue case $\epsilon_{\ell-1}\sqrt{\sigma} \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}$ for $\ell = 2$ and case $\epsilon_{\ell-2} \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}\sqrt{\sigma}$ for $\ell = 3$ separately.

- $\epsilon_{\ell-1}\sigma \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}$ for $\ell = 2$(i.e. $\epsilon_1\sigma \le \mu_1 - \mu_2 \le \epsilon_0$): We know agents in level at least 2 will observe at least $N_{K,a}^{\mathsf{FDP}} T_1/2$ pulls of arm $a$ for $a \in \{1,2\}$. By $W_2$, for any

agent in level at least 2, we have

$$|\bar{\mu}_a^t - \mu_a| \le \sqrt{\frac{3\log(T)}{\sigma N_{K,a}^{\mathsf{FDP}} T_1/2}}.$$

By Assumption 3.1, we have

$$\begin{aligned}
\hat{\mu}_1^t - \hat{\mu}_2^t &\ge \bar{\mu}_1^t - \bar{\mu}_2^t - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&\ge \mu_1 - \mu_2 - \sqrt{\frac{3\log(T)}{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \sqrt{\frac{3\log(T)}{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&\ge \frac{\sqrt{\sigma}}{4\sqrt{N_{K,1}^{\mathsf{FDP}} T_1}} + \frac{\sqrt{\sigma}}{4\sqrt{N_{K,2}^{\mathsf{FDP}} T_1}} - \sqrt{\frac{3\log(T)}{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} \\
&\quad - \sqrt{\frac{3\log(T)}{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&> 0.
\end{aligned}$$

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

- $\epsilon_{\ell-1}\sigma \le \mu_1 - \mu_2 \le \epsilon_{\ell-2}$ for $\ell > 2$: By claim A.9, for any agent $t$ in level at least $\ell$, that agent will observe at least $T_{\ell-1}$ arm $a$ pulls. By $W_2$, we have

$$|\bar{\mu}_a^t - \mu_a| \le \sqrt{\frac{3\log(T)}{T_{\ell-1}}}.$$

By Assumption 3.1, we have

$$\begin{aligned}
\hat{\mu}_1^t - \hat{\mu}_2^t &\ge \bar{\mu}_1^t - \bar{\mu}_2^t - \frac{2C_{\mathrm{est}}}{\sqrt{T_{\ell-1}}} \\
&\ge \mu_1 - \mu_2 - 2\sqrt{\frac{3\log(T)}{T_{\ell-1}}} - \frac{2C_{\mathrm{est}}}{\sqrt{T_{\ell-1}}} \\
&\ge \sqrt{\frac{\sigma}{16T_{\ell-1}}} - 2\sqrt{\frac{3\log(T)}{T_{\ell-1}}} - \frac{2C_{\mathrm{est}}}{\sqrt{T_{\ell-1}}} \\
&> 0.
\end{aligned}$$

Therefore agents in level at least $\ell$ will all pull arm 1.

- $\epsilon_{\ell-2} < \mu_1 - \mu_2 < \epsilon_{\ell-2}\sigma$ for $\ell = 3$ (i.e. $\epsilon_1 < \mu_1 - \mu_2 < \epsilon_1\sigma$): By Claim A.9, for any agent $t$ in level at least 3, that agent will observe at least $T_1 N_{K,a}^{\mathsf{FDP}} \sigma^2/2$ arm $a$ pulls (just from

the first level). By $W_2$, we have

$$|\bar{\mu}_a^t - \mu_a| \le \sqrt{\frac{3\log(T)}{\sigma^2 N_{K,a}^{\mathsf{FDP}} T_1/2}}.$$

By Assumption 3.1, we have

$$\begin{aligned}
\hat{\mu}_1^t - \hat{\mu}_2^t &\ge \bar{\mu}_1^t - \bar{\mu}_2^t - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&\ge \mu_1 - \mu_2 - \sqrt{\frac{3\log(T)}{\sigma^2 N_{K,1}^{\mathsf{FDP}} T_1/2}} - \sqrt{\frac{3\log(T)}{\sigma^2 N_{K,2}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&\ge \frac{1}{4\sqrt{\sigma N_{K,1}^{\mathsf{FDP}} T_1}} + \frac{1}{4\sqrt{\sigma N_{K,2}^{\mathsf{FDP}} T_1}} - \sqrt{\frac{3\log(T)}{\sigma^2 N_{K,1}^{\mathsf{FDP}} T_1/2}} \\
&\quad - \sqrt{\frac{3\log(T)}{\sigma^2 N_{K,2}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,1}^{\mathsf{FDP}} T_1/2}} - \frac{C_{\mathrm{est}}}{\sqrt{\sigma^2 N_{K,2}^{\mathsf{FDP}} T_1/2}} \\
&> 0.
\end{aligned}$$

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

- $\epsilon_{\ell-2} < \mu_1 - \mu_2 < \epsilon_{\ell-2}\sigma$ for $\ell > 3$: Since $\mu_1 - \mu_2 < \epsilon_{\ell-2}\sigma < \epsilon_{\ell-3}$, by Claim A.9, for any agent $t$ in level at least $\ell$, that agent will observe at least $T_{\ell-2}\sigma^2$ arm $a$ pulls (just from level $\ell - 2$). By $W_2$, we have

$$|\bar{\mu}_a^t - \mu_a| \le \sqrt{\frac{3\log(T)}{\sigma^2 T_{\ell-2}}}.$$

By Assumption 3.1, we have

$$\begin{aligned}
\hat{\mu}_1^t - \hat{\mu}_2^t &\ge \bar{\mu}_1^t - \bar{\mu}_2^t - \frac{2C_{\mathrm{est}}}{\sqrt{\sigma^2 T_{\ell-2}}} \\
&\ge \mu_1 - \mu_2 - 2\sqrt{\frac{3\log(T)}{\sigma^2 T_{\ell-2}}} - \frac{2C_{\mathrm{est}}}{\sqrt{\sigma^2 T_{\ell-2}}} \\
&\ge \frac{1}{4\sqrt{\sigma T_{\ell-2}}} - 2\sqrt{\frac{3\log(T)}{T_{\ell-1}}} - \frac{2C_{\mathrm{est}}}{\sqrt{T_{\ell-1}}} \\
&> 0.
\end{aligned}$$

Therefore agents in level at least $\ell$ will all pull arm 1. □

### A.5.2 Finishing the proof of Theorem 6.1 for $K = 2$ arms

We set the parameters $T_\ell$ for each level $\ell \in \{1, \dots, L-1\}$:

$$T_\ell = T^{\gamma_\ell}/\sigma^3, \quad \text{where} \quad \gamma_\ell := \frac{2^{L-1} + 2^{L-2} + \cdots + 2^{L-\ell}}{2^{L-1} + 2^{L-2} + \cdots + 1} = \frac{2^L - 2^{L-\ell}}{2^L - 1}. \tag{8}$$

Therefore, $T_\ell / T_{\ell-1} \geq T^{1/2^L} \geq \sigma^4$, as required by Eq. (7).
The $L$-th layer comprises all remaining nodes, hence

$$T_L = (T - T_1 \cdot L_K^{\mathsf{FDP}} \cdot \sigma^2 - (T_2 + \cdots + T_{\ell-1})\sigma^3)/\sigma^3. \tag{9}$$

*Proof.* By Claim A.10, the regret conditioned the intersection of clean events is at most

$$\begin{aligned}
&\max\left(T_1 L_K^{\mathsf{FDP}} \sigma^2, \max_{\ell \geq 2} \epsilon_{\ell-1}\sigma(T_1 L_K^{\mathsf{FDP}}\sigma^2 + T_2\sigma^3 + \cdots + T_\ell \sigma^3)\right) \\
\leq &\max\left(T_1 L_K^{\mathsf{FDP}} \sigma^2, \max_{\ell \geq 2} 2\epsilon_{\ell-1} T_\ell \sigma^4\right) \\
= &O\left(T^{2^{L-1}/(2^L-1)} \log^2(T)\right).
\end{aligned}$$

□

### A.5.3 Finishing the proof of Theorem 6.2 for $K = 2$ arms

We set the parameters as follows:

$$\begin{aligned}
L &= \log(T)/\log(\sigma^4), \\
T_\ell &= \sigma^{4\ell} \quad \text{for} \quad \ell \in \{1, \dots, L-1\}.
\end{aligned} \tag{10}$$

$T_L$ is defined via Eq. (9). Note that these settings satisfy Eq. (7), as required.

*Proof.* Recall from Appendix A.5.1 that $\epsilon_\ell = \Theta(1/\sqrt{T_\ell \sigma})$ for $\ell \in [L-1]$ and $\epsilon_0 = 1$.
Consider two cases:

- $\Delta < \epsilon_{L-1}\sigma$. In this case, notice that even always picking the sub-optimal arm gives expected regret at most $T(\mu_1 - \mu_2) = T\Delta = O(T^{1/2}\,\mathrm{polylog}(T))$. On the other hand, $T^{1/2} = O(\mathrm{polylog}(T)/\Delta)$. So, regret is $O(\min(1/\Delta, T^{1/2})\,\mathrm{polylog}(T))$.
- $\Delta \geq \epsilon_{L-1}\sigma$. In this case, we can find $\ell \in \{2, ..., L\}$ such that $\epsilon_{\ell-1}\sigma \leq \Delta < \epsilon_{\ell-2}\sigma$. By Claim A.10, we can upper bound the regret by

$$\begin{aligned}
&\Delta \cdot (T_1 L_K^{\mathsf{FDP}} \sigma^2 + T_2 \sigma^3 + \cdots T_{\ell-1}\sigma^3) \\
= &O(\Delta T_{\ell-1}\sigma^3) \\
= &O(\Delta T_{\ell-2}\sigma^7) \\
= &O(\Delta \cdot \frac{1}{\epsilon_{\ell-2}^2} \cdot \sigma^6) \\
= &O(\Delta \cdot \frac{1}{\Delta^2} \cdot \sigma^8) \\
= &O(\mathrm{polylog}(T)/\Delta).
\end{aligned}$$

We also have $1/\Delta \le 1/(\epsilon_{L-1}\sigma) = O(T^{1/2})$. So, regret is $O(\min(1/\Delta, T^{1/2})\ \mathrm{polylog}(T))$.

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

### A.5.4 Extending the analysis to $K > 2$ arms.

Here we discuss how to extend Theorems 6.2 and 6.2 to $K > 2$ arms. The analysis is very similar to the $K = 2$ case, so we only sketch the necessary changes.

*Proof Sketch.* We still wlog assume arm 1 has the highest mean (i.e. $\mu_1 \ge \mu_a, \forall a \in \mathcal{A}$. We first extend the clean events (i.e. $W_1, W_2, W_3, W_4$) in Appendix A.5.1 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 $W_2 = \bigcap_{t,a,\tau_1,\tau_2} W_2^{t,a,\tau_1,\tau_2}$. The difference is that now $a$ is taken over $K$ arms instead of 2 arms. For $W_3$, we change the definition $W_3^{\ell,a} = \bigcup_u \left( W_3^{\ell,u,a,high} \cap \left( \bigcap_{a' \ne a} W_3^{\ell,u,a',low} \right) \right)$ and $W_3 = \bigcap_{\ell,a} W_3^{\ell,a}$. We extend $W_4$ in a similar way: define $W_4^a$ as $\bigcup_u \left( W_4^{u,a,high} \cap \left( \bigcap_{a' \ne a} W_4^{u,a',low} \right) \right)$ 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 A.9 and Claim A.10. The statement of Claim A.9 should be changed to "For any arm $a$ and $2 \le \ell \le L$, if $\mu_1 - \mu_a \le \epsilon_{\ell-1}$, then for any $u \in [\sigma]$, there are at least $T_\ell$ pulls of arm $a$ in groups $G_{\ell,u,1}, G_{\ell,u,2}, ..., G_{\ell,u,\sigma}$ and there are at least $T_\ell\sigma(\sigma-1)$ pulls of arm $a$ in the $\ell$-th level $\Gamma$-groups". The statement of Claim A.10 should be changed to "For any $2 \le \ell \le L$, if $\epsilon_{\ell-1}\sigma \le \mu_1 - \mu_a < \epsilon_{\ell-2}\sigma$, there are no pulls of arm $a$ in groups with level $\ell, ..., L$."

The proof of Claim A.10 can be easily changed to prove the new version by changing "arm 2" to "arm $a$". The proof of Claim A.9 needs some additional argument. In the proof of Claim A.9, we show that $\hat{\mu}_a^t - \hat{\mu}_{3-a} > 0$ for agent $t$ in the chosen groups. When extending to more than 2 arms, we need to show $\hat{\mu}_a^t - \hat{\mu}_{a'}^t > 0$ for all arm $a' \ne a$. The proof of Claim A.9 goes through if $\mu_1 - \mu_{a'} \le \epsilon_{\ell-2}$ since then there will be enough arm $a'$ pulls in level $\ell-1$. We need some additional argument for the case when $\mu_1 - \mu_{a'} > \epsilon_{\ell-2}$. Since $\mu_1 - \mu_{a'} > \epsilon_{\ell-2} > \epsilon_{\ell-1}\sigma$, we can use the same proof of Claim A.10 (which rely on Claim A.9 but for smaller $\ell$'s) to show that there are no arm $a'$ pulls in level $\ell$ and therefore $\hat{\mu}_a^t - \hat{\mu}_{a'}^t > 0$.

Finally we proceed to bound the regret conditioned on the intersection of clean events happens. The analysis for $K = 2$ bounds 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. □