Incentivizing Exploration with Selective Data Disclosure*

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

First version: November 2018
This version: February 2020

arXiv:1811.06026v3 [cs.GT] 19 Feb 2020

Abstract

We study the design of rating systems that incentivize (more) efficient social learning among self-interested agents. Agents arrive sequentially and are presented with a set of possible actions, each of which yields a positive reward with an unknown probability. A disclosure policy sends messages about the rewards of previously-chosen actions to arriving agents. These messages can alter agents' incentives towards *exploration*, taking potentially sub-optimal actions for the sake of learning more about their rewards. Prior work achieves much progress with disclosure policies that merely recommend an action to each user, but relies heavily on standard, yet very strong rationality assumptions.

We study a particular class of disclosure policies that use messages, called unbiased subhistories, consisting of the actions and rewards from a subsequence of past agents. Each subsequence is chosen ahead of time, according to a predetermined partial order on the rounds. We posit a flexible model of frequentist agent response, which we argue is plausible for this class of "order-based" disclosure policies. We measure the success of a policy by its *regret*, i.e., the difference, over all rounds, between the expected reward of the best action and the reward induced by the policy. A disclosure policy that reveals full history in each round risks inducing herding behavior among the agents, and typically has regret linear in the time horizon T. Our main result is an order-based disclosure policy that obtains regret $\tilde{O}(\sqrt{T})$. This regret is known to be optimal in the worst case over reward distributions, even absent incentives. We also exhibit simpler order-based policies with higher, but still sublinear, regret. These policies can be interpreted as dividing a sublinear number of agents into constant-sized focus groups, whose histories are then revealed to future agents.

*Earlier versions of this manuscript have been titled "Incentivizing Exploration with Unbiased Histories".

†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.
¶University of Minnesota, Minneapolis, MN. Email: zsw@umn.edu. Research done during a postdoc at MSR-NYC.

Contents

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 customers. 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 provide each consumer with an optimal experience by presenting the most comprehensive and comprehensible information. On the other hand, platforms need to encourage consumers to explore infrequently-selected choices in order to learn more about them. The said exploration, while beneficial for the common good, is often misaligned with incentives of individual consumers, who are often reluctant to explore and prefer to err on the side of less risk. This tension between exploration and incentives lies at the heart of our study. We resolve it by providing a platform design that achieves a near-optimal rate of social learning and withholds little information from consumers. Our platform design presents each consumer with a subset of ratings, and allows the consumer to draw her own conclusions. Each user's subset is determined in advance, and so cannot be biased to make a particular action look good; we call it an *unbiased subhistory*.

As an example, consider a diner searching for restaurants on a platform like Yelp. If all restaurants are displayed alongside their past average rating, the diner is likely to choose the restaurant with the highest rating, especially if that restaurant has a large number of reviews. But then future diners will follow in her footsteps, and the platform never learns about alternative restaurants. If the platform instead recommends only a single restaurant with no accompanying data in an attempt to nudge the diner towards that choice, the diner may lose faith in the system and abandon the platform altogether.

Our design presents a third option: the platform displays a subset of past ratings, chosen in advance of deployment. Thus, the diner might see, *e.g.*, every fifth review entered into the platform, or the first hundred reviews, etc. As different diners see different subhistories, random variation incentivizes them to make different choices. If the subhistories are intertwined in a certain way, we prove that the diners eventually receive enough information to make *optimal* dining choices. As a result, the average regret of a diner, measured as the difference in reward between her selected restaurant and the best one, is vanishing as the population grows. In fact, we match the best possible regret rates for the version without incentives.

Background and scope. Absent incentive constraints, the so-called *exploration-exploitation tradeoff* has received much attention over the past decades, mainly in the relatively simple abstraction known as *multi-armed bandits* (see books [21, 15, 47, 34] for background). In this literature, the social planner repeatedly selects from a set of actions, 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. By coordinating actions across time, the planner can guarantee an average reward which converges to that of the optimal action in hindsight with regret rate proportional to the square root of the time horizon.

In the rating platform setting, the actions are not chosen by a social planner, but rather a sequence of self-interested customers, henceforth *agents*. An agent's incentives are skewed towards exploitation: indeed, the costs of exploration are all hers, whereas the benefits are spread over many. In particular, an agent that shows up only once will only exploit, as the benefits of exploration are only accrued by future agents. This behavior can cause *herding*, in which all agents eventually take a sub-optimal action, as it yields maximum expected payoff given the available information. In particular, if the platform adopts the *full-disclosure policy*, whereby each agent sees the full history of observations from the previous agents, then herding happens with constant probability (*e.g.*, see Chapter 11.2 in [47]). Extreme behaviors aside, some actions may get explored at a very suboptimal rate, and may suffer from selection bias (*e.g.*, a movie is only seen and reviewed by the fans).

This situation can be circumvented by a rating system that induces agents to take explorative actions, an idea called *incentivizing exploration*. One way rating platforms induce exploration is via payments. For example, a platform like Yelp might offer coupons to diners for trying certain restaurants. However, exploration-inducing payments introduce selection bias, and are often financially or technologically infeasible. An alternative that we explore here is to rely on information asymmetry. The rating system can choose a *disclosure policy* to selectively release information about the past actions and rewards. The agent then chooses an action, using the available information as input. The problem of incentivizing exploration via information asymmetry was introduced in [32] and further explored in subsequent work [16, 38, 39, 7, 13, 8]. Incentivizing exploration via payments has been studied in [19, 26, 17].

All this work relies heavily on the standard assumptions of Bayesian rationality and the "power to commit" (*i.e.*, that users trust that the platform actually implements the policy that it claims to implement). However, these assumptions appear quite problematic in the context of ratings systems of actual online markets such as those mentioned above. In particular, much of the prior work suggests policies that merely recommend an action to each agent, without any other supporting information, and moreover recommend exploratory actions to some randomly selected users. This works out extremely well in theory. However, realistic users may hesitate to follow such a policy – because of limited rationality, insufficient trust in the platform, aversion to exploration, or preference for detailed and interpretable information (or all of the above).

Our model. We strive to design information disclosure policies which mitigate these issues by taking a more realistic view of user behavior while (still) striking a good balance between exploration and exploitation. While making some assumptions on user behavior appears unavoidable, we relax the assumptions in two crucial ways:

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

2. We allow a flexible frequentist choice model: an agent can choose an action in any way that is consistent with (a slightly narrower version of) her confidence intervals.

Let us elaborate on these two tenets. The full-disclosure policy is arguably most intelligible and satisfactory for the users. It is also relatively easy to audit, which should increase the platform's "power to commit" to this policy. Hence, any assumptions on user behavior are most plausible if we only need to make them against the full-disclosure policy. Now, 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. Essentially, we allow the agent to pick an arbitrary estimate in each confidence interval, and choose an action with the highest estimate.[1] A more detailed discussion of the economic assumptions can be found later in the Introduction.

Our information disclosure policy proceeds as follows. A partial order on the rounds is carefully constructed and fixed throughout (and can be made public w.l.o.g.). An agent arriving in round t sees full history for a fixed subset S_t of rounds, namely for all rounds s that precede t in the partial order. No other information is revealed. In particular, unlike all prior work on incentivized exploration, we do not make explicit recommendations for which action to choose next. Such disclosure policies are called *order-based*. 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 for this mechanism.

Put differently, we construct a unidirectional communication network for the agents, and let them engage in social learning with full disclosure (*i.e.*, each agent communicates its full information set). We construct the network so as to ensure that full disclosure does not lead to *herding*, and instead results in a near-optimal exploration-exploitation balance.

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

Our results and techniques. We have arrived at a concrete mathematical problem: design an order-based disclosure policy so as to optimize regret. We focus on the fundamental case of a constant number of actions. Our main result is policy that attains near-optimal $\tilde{O}(\sqrt{T})$ regret rate. This policy also obtains the optimal instance-dependent regret rate $\tilde{O}(\frac{1}{\Delta})$ for problem instances with gap Δ, without knowing the Δ in advance. In particular, we match the best possible regret rates for the multi-armed bandit problem. Our disclosure policy also ensures a desirable property that each agent t sees a substantial fraction of history available at time t: namely, our policy reveals a subhistory of size at least $\Omega(t/\text{polylog}(T))$.

The main challenge is that the agents still follow exploitation-only behavior, just like they do for the full-disclosure policy, albeit based only on a portion of history. Recall that our disclosure policy controls the flow of information (*i.e.*, who sees what), but not its *content*.

The first step is to obtain any substantial improvement over the full-disclosure policy. We accomplish this with a relatively simple policy which runs the full-disclosure policy "in parallel" on several disjoint subsets of agents, collects all data from these runs and discloses it to all remaining agents. In practice, these subsets may correspond to multiple "focus groups". While any single run of the full-disclosure policy may get stuck on a suboptimal arm, having these parallel runs ensure that sufficiently many of them will "get lucky" and provide some exploration. This simple policy achieves $\tilde{O}(T^{2/3})$ regret. Conceptually, it implements a basic bandit algorithm that explores

[1] More precisely, we restrict the estimates to the central portion of each confidence interval, which accounts for a constant fraction of the interval's width.

uniformly for a pre-set number of rounds, then picks one arm for exploitation and stays with it for the remaining rounds. We think of this policy as having two "levels": Level 1 contains the parallel runs, and Level 2 is everything else.

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

Discussion: economic aspects. We argue that order-based information disclosure policies require substantially weaker trust and rationality assumptions compared to information disclosure policies in prior work on incentivized exploration. The latter are bandit algorithms which recommend an action to each agent, under a Bayesian incentive-compatibility condition to ensure that the agents follow recommendations (and strong implicit assumptions of trust and rationality). We will refer to this work as *Bayesian incentivized exploration*. Several distinct issues are in play:

- *Whether agents understand the announced policy.* We only need an agent to understand that he is given some unbiased history. It does not matter what is the subset of rounds 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.
- *Whether agents trust the principal's intent 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., that arms' average rewards are within the statistical deviations). This should create incentives for the principal not to manipulate the policy. Note that similar checks appear virtually impossible for bandit algorithms.
- *Whether agents trust the principal to implement the stated policy without bugs.* Faithfully revealing a subhistory is arguably easy, whereas (as noted above) debugging an actual bandit algorithm in a large-scale production-level system tends to be quite complicated.
- *Whether agents react according to our choice model.* Our framework encourages an agent to interpret the subhistory as just a set of data points collected by an algorithm. In particular, 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 which data points are included in the subhistory), or what was the data seen by an agent when she chose her action (because all that data is included in the subhistory). The system can provide summary statistics, so that agents would not need to actually look at the raw data.[2] Whereas verifying that a recommendation policy is incentive-compatible typically requires a sophisticated Bayesian reasoning.

[2]Trusting the summary statistics is a relatively minor issue, one where "cheating" can, in principle, be easily verified.

Would a detail-oriented user be happy with an unbiased subhistory which contains only a fraction of the full data? Such subhistory still contains a large number of observations, probably more than a typical user ever needs. The platform pre-selects the observations in an unbiased way, which is arguably not undersirable and possibly useful for the user. Note that even a small subhistory gives much more details than a recommendation-only policy.

Our choice model allows several deviations from rationality. First, we allows for a considerable amount of optimism/pessimism, a.k.a. risk preferences. An optimistic (risk-loving) agent may estimate each action's expected reward as a value towards the top of the corresponding confidence interval, and pessimistic (risk-averse) estimates would be skewed towards the bottom. Second, we allow Softmax-like choices that randomize around the best actions. Indeed, we allow each reward estimate to be randomized, as long as it falls in the corresponding interval. Third, we allow agents to have strong initial beliefs. This is because we make no assumptions on the reward estimates after seeing $n \in [1, N_{\mathtt{est}}]$ samples, for some constant $N_{\mathtt{est}}$ in the model. Eventually, the effect of initial beliefs is drowned out.

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 algorithm knows the prior, or whether users have correct beliefs, or whether they can handle the congnitive load of Bayesian reasoning. While our model does not rely on Bayesian foundations, it *is* consistent with a version of Bayesian rationality. Specifically, it is consistent with Bayesian-rational choices based on independent Beta-Bernoulli beliefs (although these beliefs are inconsistent with our rewards model), see Remark 2.4 for details.

For the sake of the analysis, we assume that the mean rewards lie in the $[1/3, 2/3]$ interval,[3] whereas the agents' response model is oblivious to this. This makes sense for two reasons. First, agents may operate under incomplete information and be unaware of this restriction. Second, typical users of a recommendation system are unsophisticated. They are more likely to follow empirical averages rather than be fully rational (whether Bayesian or not). In particular, a realistic user would not estimate mean rewards as (at least) $1/3$ after observing a long sequence of 0 rewards. Alternatively, we can project all reward estimates into the $[1/3, 2/3]$ interval and assume random tie-breaking. Either variant works for the rest of the paper, see Remark 2.3.

Discussion: regret rates. Let us compare the regret rates in prior work on Bayesian incentivized exploration against ours. The prior work [38] also achieves the optimal regret rate $O(\min(\sqrt{T}, \frac{1}{\Delta}\log T))$ for a constant number of actions, where Δ is the "gap". The $O()$ in this work includes a "constant" that can get arbitrarily large depending on the Bayesian prior, whereas the $O()$ in our paper hides the dependence on a constant from our choice model.

Most prior work either assumes $K = 2$ actions (*e.g.*, [32, 16, 13, 7]), or targets the case of constant K (*e.g.*, [38, 39]). The regret bounds in prior work, as well as ours, scale exponentially in K. This dependence is grossly suboptimal for multi-armed bandit algorithms without incentives, where one can achieve regret rates that scale as $\sqrt{K}$. A very recent, yet unpiblished manuscript [45] achieves Bayesian incentivized exploration with poly(K) regret scaling, albeit only for independent priors and only for Bayesian regret (*i.e.*, regret in expectation over the Bayesian prior).

[3]This interval can be replaced with $[\epsilon, 1-\epsilon]$ for any absolute constant $\epsilon > 0$ (which then propagates throughout).

1.1 Related work

The problem of incentivizing exploration via information asymmetry was introduced in [32], under Bayesian rationality and (implicit) power-to-commit assumptions. Information disclosure policies are unrestricted, and therefore they can be reduced to recommendation policies w.l.o.g., by a version of *revelation principle*. The original problem (which corresponds to stochastic K-armed bandits with a Bayesian prior) was largely resolved in [32] and the subsequent work [38, 39, 45]. The technical results come in a variety of flavors, concerning regret rates [32, 38, 45], a black-box reduction from arbitrary bandit algorithms to incentive-compatible ones [38], Bayesian-optimal policies for special cases [32], and exploring all "explorable" actions [39]. Several extensions were considered: to contextual bandits [38], repeated games and misaligned incentives [39], and social networks [7]. Several other papers study related, but technically different models: a version with time-discounted utilities [13]; a version with monetary incentives [19, 17]; a version with a continuous information flow and a continuum of agents [16]; coordination of costly "exploration decisions" when they are separate from "payoff-generating decisions" [30, 35, 36].

Incentivized exploration is closely related to two prominent subareas of theoretical economics: *information design* and *social learning*. Information design [10, 48] studies the design of information disclosure policies and incentives that they create. In particular, a single round of incentivized exploration is a version of the *Bayesian Persuasion* game [25], 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 [23, 31, 37, 40], including conditions that make assumptions on the agent behavior [41]. A large literature on *social learning* studies long-lived agents that learn in a shared environment, with no principal to coordinate them. Most of this literature posits that only actions of the past agents (but not their outcomes) are observable in the future. A prominent topic is the presence or absence of herding phenomena. Scenarios with long-lived learning agents that observe both actions and rewards of one another have been studied in [14, 28].

Full-disclosure policy, and closely related "greedy" (exploitation-only) algorithm in multi-armed bandits, have been a subject of a recent line of work [44, 27, 9, 43]. A common theme is that the greedy algorithm performs well in theory, under substantial assumptions on heterogeneity of the agents. Yet, it suffers $\Omega(T)$ regret in the worst case (see Chapter 11.2 in [47]).

Exploration-exploitation problems with incentives issues naturally arise in a variety of scenarios, *e.g.*, dynamic pricing [29, 12, 6], dynamic auctions [1, 11, 24], pay-per-click ad auctions [5, 18, 4], and human computation [22, 20, 46]. For more background, see Chapter 11 of [47].

2 Model and Preliminaries

We study the multi-armed bandit problem in a social learning context, in which a principal faces a sequence of T myopic agents. There is a set $\mathcal{A}$ of K possible actions, a.k.a. *arms*. At each round $t \in [T]$, 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 $r_t \in \{0,1\}$ that is immediately observed by the principal. The reward from pulling an arm $a \in \mathcal{A}$ is drawn independently from Bernoulli distribution $\mathcal{D}_a$ with an unknown mean μ_a. An agent does not observe anything from the previous rounds, other than the message m_t. The problem instance is defined by (known) parameters K, T and the (unknown) tuple of mean rewards, $(\mu_a : a \in \mathcal{A})$. We are interested in *regret*, defined as

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

(The expectation is over the chosen arms a_t, which depend on randomness in rewards, and possibly in the algorithm.) The principal chooses messages m_t according to an online algorithm called *disclosure policy*, with a goal to minimize regret. We assume that mean rewards are bounded away from 0 and 1, to ensure sufficient entropy in rewards. For concreteness, we posit $\mu_a \in [\frac{1}{3}, \frac{2}{3}]$.

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

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

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

We focus on disclosure policies of a particular form, 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 message is called *unbiased subhistory*. 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*.

Note that order-based disclosure policies are *transitive*, in the following sense:

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

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 estimator. (Ties are broken according to an arbitrary rule that is the same for all agents.) The basic model is that $\hat{\mu}_{t,a}$ is the sample average for arm a over the subhistory m_t, as long as it includes at least one sample for a; else, $\hat{\mu}_{t,a} \geq \frac{1}{3}$.

We allow a much more permissive model that allows agents to 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).

Assumption 2.1. *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 $N_{\mathrm{est}} \in \mathbb{N}$ and $C_{\mathrm{est}} = \frac{1}{16}$, and for all agents $t \in [T]$ and arms $a \in \mathcal{A}$ it holds that*

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

Also, $\hat{\mu}_a^t \geq \frac{1}{3}$ if $N_{t,a} = 0$. (NB: we make no assumption if $1 \leq N_{t,a} < N_{\mathrm{est}}$.)

Remark 2.2. In the $\hat{\mu}_a^t \geq \frac{1}{3}$ assumption above, the $\frac{1}{3}$ can be replaced with an arbitrary strictly positive constant, with very minor changes in the proofs. In words, we posit that each agent's initial belief on each arm is bounded away from zero.

Remark 2.3. The choice model in Assumption 2.1 is oblivious to the fact that the mean rewards are restricted to lie in the $[1/3, 2/3]$ interval. As pointed out in the Introduction, we could project all reward estimates into this interval (so that an estimate smaller than $1/3$ becomes exactly $1/3$, and similarly for the upper bound), and assume random tie-breaking. This variant works, too, with minimal changes to the analysis.

Remark 2.4. While our frequentist choice model does not rely on Bayesian foundations, we note that it *is* consistent with a version of Bayesian rationality. Indeed, suppose an agent has an independent Beta-Bernoulli prior for each arm a, and the estimate $\hat{\mu}_a^t$ is the posterior mean reward given the history. Then the estimates satisfy Assumption 2.1 for a large enough constant N_{est} which depends on the priors. (This is because for Beta-Bernoulli priors the absolute difference between the posterior mean and the empirical mean scales as 1/#samples.) However, such beliefs would necessarily be *inconsistent* our model of rewards, as they place positive probability outside of the $[1/3, 2/3]$ interval.

Assumption 2.5. *In each round t, the estimates $\hat{\mu}_{t,a}$ 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 an* estimate function f_t *from anonymized subhistories to $[0,1]^K$, so that the estimate vector $(\hat{\mu}_{t,a} : a \in \mathcal{A})$ equals $f_t(m'_t)$. This function is drawn from some fixed distribution over estimate functions.*

Connection to multi-armed bandits. The special case when each message m_t is an arm, and the t-th agent always chooses this arm, corresponds to a standard multi-armed bandit problem with IID rewards. Thus, regret in our problem can be directly compared to regret in the bandit problem with the same mean rewards $(\mu_a : a \in \mathcal{A})$. Following the literature on bandits, we define the *gap parameter* Δ as the difference between the largest and second largest mean rewards.[4] The gap parameter is not known to the principal (in our problem), or to the algorithm (in the bandit problem). Optimal regret rates for bandits with IID rewards are as follows [2, 3, 33]:

$$\text{Reg}(T) \leq O\left(\min\left(\sqrt{KT\log T},\; \tfrac{1}{\Delta}\log T\right)\right). \tag{3}$$

This regret bound can only be achieved using *adaptive exploration*: *i.e.*, when exploration schedule is adapted to the observations. A simple example of *non*-adaptive exploration is the *explore-then-exploit* algorithm which samples arms uniformly at random for the first N rounds, for some pre-set number N, then chooses one arm and sticks with it till the end. More generally, *exploration-separating* 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. Any such algorithm suffers from $\Omega(T^{2/3})$ regret in the worst case.[5]

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

Throughout the paper, we use the standard concentration and anti-concentration inequalities: respectively, Chernoff Bounds and Berry-Esseen Theorem. The former states that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

[4] Formally, the second-largest mean reward is $\max_{a\in\mathcal{A}:\mu(a)<\mu^*} \mu(a)$, where $\mu^* = \max_{a\in\mathcal{A}} \mu(a)$.

[5] The first explicit reference we know of is [5, 18], but this fact has been known in the community for much longer.

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 2.6. *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*

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

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

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

3 Warm-up: full-disclosure paths

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

Definition 3.1. 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 constant probability, a full-disclosure path of constant length suffices to sample each arm at least once. We will build on this fact throughout.

Lemma 3.2. *There exist numbers $L_K^{\mathsf{FDP}} > 0$ and $p_K^{\mathsf{FDP}} > 0$ 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 $|S| \geq L_K^{\mathsf{FDP}}$. Under Assumption 2.1, with probability at least p_K^{FDP}, subhistory $\mathcal{H}_S$ contains at least once sample of each arm a.*

Proof. Fix any arm a. Let $L_K^{\mathsf{FDP}} = (K-1)\cdot N_{\mathtt{est}} + 1$ and $p_K^{\mathsf{FDP}} = (1/3)^{L_K^{\mathsf{FDP}}}$. We will condition on the event that all the realized rewards in L_K^{FDP} rounds are 0, which occurs with probability at least p_K^{FDP} under Assumption 2.1. 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 $N_{\mathtt{est}}+1$ rounds. Let t be the round in which arm a' is pulled exactly $N_{\mathtt{est}}+1$ times. By Assumption 2.1, we know

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

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. □

We provide a simple disclosure policy based on full-disclosure paths. The policy follows the "explore-then-exploit" paradigm. The "exploration phase" comprises the first $N = T_1 \cdot L_K^{\mathsf{FDP}}$ rounds, and consists of T_1 full-disclosure paths of length L_K^{FDP} each, where T_1 is a parameter. In the "exploitation phase", each agent $t > N$ receives the full subhistory from exploration, *i.e.*, $m_t = \mathcal{H}_{[N]}$. The info-graph for this disclosure policy is shown in Figure 1.

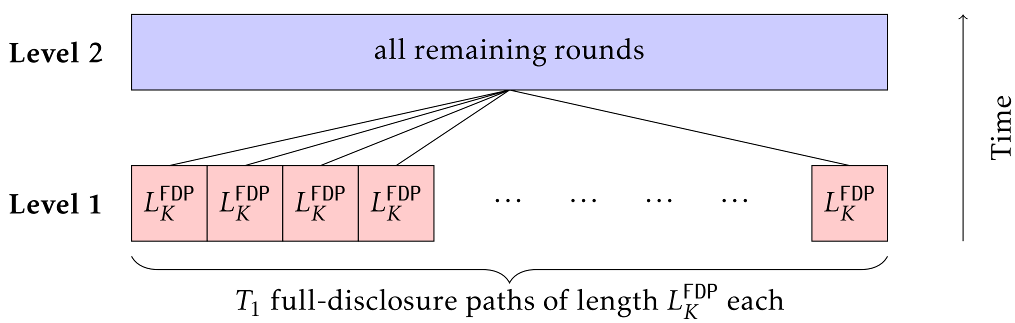

Figure 1: Info-graph for the 2-level policy.

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

Theorem 3.3. *The two-level policy with parameter $T_1 = T^{2/3}\log(T)^{1/3}$ achieves regret*

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

Remark 3.4. 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 Δ is known to the principal, then (for an appropriate tuning of parameter T_1) we can achieve regret $\mathrm{Reg}(T) \leq O_K(\log(T)\cdot\Delta^{-2})$.

The proof can be found in Section 3.1. 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^{FDP} (*i.e.*, present in the subhistory $\mathcal{H}_S$. Indeed, this number, denoted $N_{K,a}^{\mathsf{FDP}}$, is the same for all such paths. Then,

Lemma 3.5. *Suppose the info-graph contains T_1 full-disclosure paths of L_K^{FDP} rounds each. Let N_a be the number of samples of arm a collected by all paths. Then with probability at least $1-\delta$,*

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

3.1 Detailed analysis: proof of Theorem 3.3

We will set T_1 later in the proof, depending on whether the gap parameter Δ 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 3.5, 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 τ realized rewards of a coming from the prefix of τ cells in its reward tape. Define $W_2^{a,\tau}$ to be the event that the empirical mean of the first τ realized rewards in the tape of arm a is at most $\sqrt{\frac{2\log(T)}{\tau}}$ away from μ_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 2.1, we have

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

Therefore,

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

So the second-level agents will pick an arm a which has μ_a at most $6\sqrt{\frac{\log(T)}{p_K^{\mathsf{FDP}} T_1}}$ away from μ_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 $O(T^{2/3} \log(T)^{1/3})$.

4 Adaptive exploration with a three-level disclosure policy

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

For the sake of intuition, consider the framework of bandit algorithms with limited adaptivity [42]. 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. A single round of adaptivity corresponds to explore-then-exploit paradigm. Our goal here is to implement one extra adaptivity round, and this is what the middle level accomplishes.

Construction 4.1. *The* three-level policy *is 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^{FDP} each, and each agent t in the exploitation level sees full history from exploration (see Figure 2).*

The middle level consists of σ disjoint subsets of T_2 agents each, called second-level groups. *Each second-level group G has the following property:*

$$\text{all nodes in } G \text{ are connected to the same nodes outside of } G\text{, but not to one another.} \tag{4}$$

The full-disclosure paths in the first level are also split into σ 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 $\mathcal{H}_S$, where S is the set of all rounds in G.*

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

Theorem 4.2. *For two arms, the three-level policy achieves regret*

$$\mathrm{Reg}(T) \le O\left(T^{4/7} \log T\right).$$

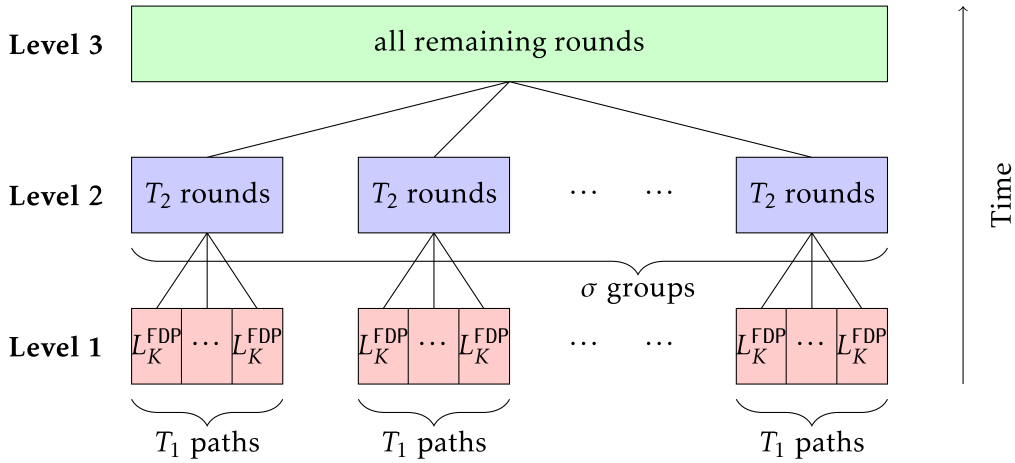

Figure 2: Info-graph for the three-level policy. Each red box in level 1 corresponds to T_1 full-disclosure paths of length L_K^{FDP} each.

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

Let us sketch the proof of this theorem; the full proof can be found in Sections 4.1 and 4.2.

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

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

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

($\mathtt{event}_3$) *Anti-concentration in Level 1:* For each arm, some first-level subgroup collects data which makes this arm look much better than its actual mean and other arms look worse than their actual means.

($\mathtt{event}_4$) *Concentration in prefix:* The empirical mean reward of each arm a concentrates around μ_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 Section 2). For example, $\mathtt{event}_2$ 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 $\mathtt{conf}(n) = \sqrt{\log(T)/n}$. In words, this is a confidence term, up to constant factors, for n independent random samples.

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

Claim 4.3 (Negligible gap)**.** *If* $\Delta \leq 3\sqrt{2}\cdot\mathtt{conf}(T_2)$ *then* $\mathrm{Reg}(T) \leq O(T^{4/7}\log^{6/7}(T))$*.*

Another simple case is when Δ is sufficiently large, so that the data collected in any first-level group suffices to determine the best arm. The proof follows from $\mathtt{event}_1$ and $\mathtt{event}_2$.

Lemma 4.4 (Large gap). *If $\Delta \geq 4\sum_{a\in\mathcal{A}} \texttt{conf}\left(N^{\mathsf{FDP}}_{K,a}\cdot T_1\right)$ 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 not only one but all first-level groups and the data collected by all first-level groups enables agents in the third level to correctly identify the best arm.

Lemma 4.5 (Medium gap). *All agents pull arm 1 in the third level, when Δ satisfies*

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

Finally, the *small gap* case, when Δ 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 σ 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 μ_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 2.1 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 4.6 (Small gap). *All agents pull arm 1 in the third level, when Δ satisfies*

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

Wrapping up: proof of Theorem 4.2. In negligible gap case, the stated regret bound holds regardless of what the algorithm does. In the large gap case, the regret only comes from the first level, so it is upper-bounded by the total number of agents in this level, which is $\sigma\cdot L^{\mathsf{FDP}}_K\cdot T_1 = O(T^{4/7}\log T)$. In both intermediate cases, it suffices to bound the regret from the first and second levels, so

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

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

4.1 High-probability events

The following lemmas can be derived from combining Lemma 3.5 and union bound.

Lemma 4.7 (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*

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

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

Proof of Lemma 4.7. 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 $N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$ and $N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$. By Lemma 3.5

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

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 $\mathcal{T}_{s,a}^1$ 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 4.8 (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,τ_1,τ_2} be the event that the mean among the cells indexed by $\tau_1, (\tau_1+1), \ldots, \tau_2$ in the tape $\mathcal{T}_{a,s}^1$ is at most $\sqrt{\frac{2\log(T)}{\tau_2 - \tau_1 + 1}}$ away from μ_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 4.8. 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 μ_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 4.9 (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 $N_{K,a}^{\mathsf{FDP}} T_1$ cells in tape $\mathcal{T}_{a,s}^1$. Let $W_3^{s,a,high}$ be the event $\tilde{\mu}_{a,s} \geq \mu_a + 1/\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}$ and let $W_3^{s,a,low}$ be the event that $\tilde{\mu}_{a,s} \leq \mu_a - 1/\sqrt{N_{K,a}^{\mathsf{FDP}} T_1}$. 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 4.9. By Berry-Esseen Theorem and $\mu_a \in [1/3, 2/3]$, we have for any a,

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

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

$$\Pr[W_3] \geq 1 - 2(1-1/16)^{\sigma} \geq 1 - 2/T. \qquad \square$$

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 4.10 (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 τ arm a pulls is at most $\sqrt{\frac{2\log(T)}{\tau}}$ away from μ_a, where $N_{T,a}$ is the total number of arm a pulls by the end of T rounds.*

Proof of Lemma 4.10. 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 μ_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:

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

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.

4.2 Case Analysis

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

Lemma 4.11. *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. $\square$

Lemma 4.12. *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 4.13. *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 $\tau = N_{K,a}^{\mathsf{FDP}} T_1 - L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)} + 1, \dots, N_{K,a}^{\mathsf{FDP}} T_1 - 1$ and $W_2^{s_a,a,N_{K,a}^{\mathsf{FDP}} T_1,\tau}$ for $\tau = N_{K,a}^{\mathsf{FDP}} T_1, \dots, N_{K,a}^{\mathsf{FDP}} T_1 + L_K^{\mathsf{FDP}} \sqrt{T_1 \log(T)}$ 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 4.4 (Large gap case). Observe that for any group s in the first level, the empirical means satisfy

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

For any agent t in the s-th second-level group, by Assumption 2.1, we have

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

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 2.1, 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 4.5 (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 2.1, we have

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

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

Proof of Lemma 4.6 (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 4.13. 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 2.1, we have

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

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 2.1, 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. □

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

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

Theorem 5.1. *For any $L > 3$, there exists an L-level disclosure policy with regret*

$$
O_K\left(T^{2^{L-1}/(2^L-1)} \cdot \text{polylog}(T)\right).
$$

In particular, there exists a $O(\log\log(T))$-level recommendation policy with regret $O_K(T^{1/2}\,\text{polylog}(T))$.

Our second policy achieves an instance-dependent regret guarantee. This policy has the same info-graph structure as the first one in Theorem 5.1, 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 Δ even though the construction of the policy does not depend on Δ. In particular, this regret bound outperforms the one in Theorem 5.1 when Δ 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 5.2. *There exists an $O(\log(T)/\log\log(T))$-level policy such that for every multi-armed bandit instance with gap parameter Δ, the policy has regret*

$$
O_K(\min(1/\Delta, T^{1/2}) \cdot \text{polylog}(T)).
$$

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

Note for constant number of arms, this result matches the optimal regret rate (given in Equation (3)) for stochastic bandits, up to logarithmic factors.

Let us present the main techniques in our solution; the full proofs are deferred to Section 5.1. Similarly as Section 4, we first prove them in the case of 2 arms (Theorem 5.3 and Corollary 5.6). We then extend them to the case of constant number of arms (Theorem 5.7).

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

Interlacing connections between levels. A tempted approach to generalize the three-level policy is to build an L-level info-graph with the structure of a σ-ary tree: for every $l \in \{2, \dots, L\}$, each l-level group observes the sub-history from a disjoint set σ groups in level $(l-1)$. The disjoint sub-histories observed by all the groups in level l are independent, and under the small gap regime (similar to Lemma 4.6) it ensures that each arm a has a "lucky" l-level group of agents that only pull a. This "lucky" property is crucial for ensuring that both arms will be explored in level l.

However, in this construction, the first level will have σ^{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 5.2. 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 $\Theta(\log T)$ l-th level groups that observe disjoint sub-histories that take place in level $(l-1)$. Moreover, as long as the group size in levels lower than $(l-1)$ are substantially smaller than group size of level $l-1$, the "lucky" property does not break even if different groups in level l observe overlapping sub-history from levels $\{1, \dots, l-2\}$.

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

(i) Consider level $(l-1)$ and fix the last group index to be v, and consider the set of groups $\mathcal{G}_{l-1,v} = \{G_{l-1,i,v} \mid i \in [\sigma]\}$ (e.g. $G_{l-1,1,1}$ and $G_{l-1,2,1}$ circled in red in the Figure 3). The agents in any group of $\mathcal{G}_{l-1,v}$ observe the same sub-history. As a result, if the empirical mean of arm a is sufficiently high in their shared sub-history, then all groups in $\mathcal{G}_{l-1,v}$ will become "lucky" for a.

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

Amplifying groups for boundary cases. Recall in the three-level policy, the medium gap case (Lemma 4.5) corresponds to the case where the gap Δ is between $\Omega\left(\sqrt{1/T_1}\right)$ and $O\left(\sqrt{\log(T)/T_1}\right)$. This is a boundary case since Δ is neither large enough to conclude that with high probability

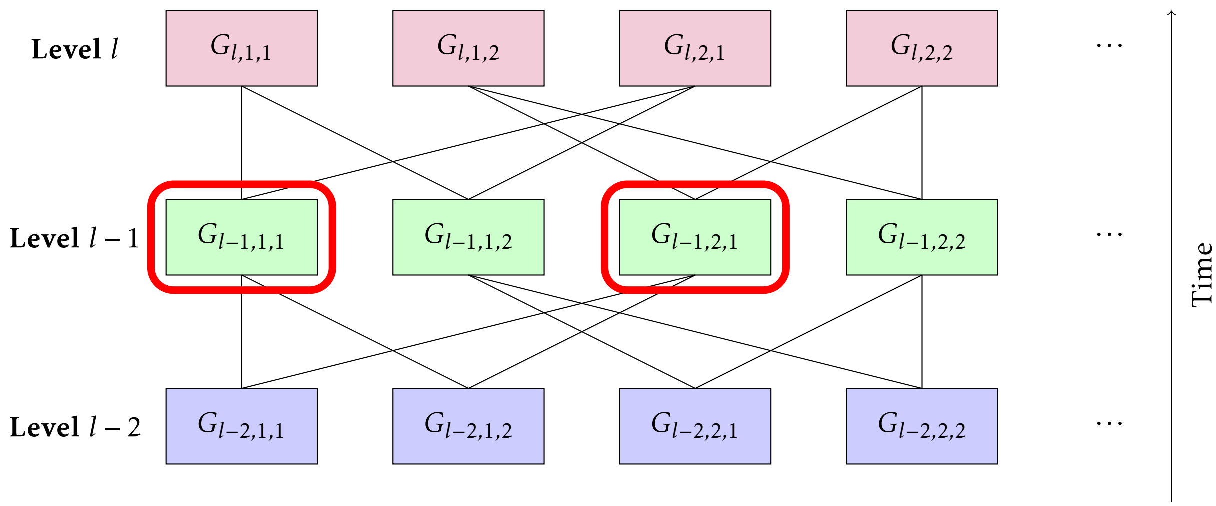

Figure 3: Interlacing connections between levels for the L-level policy.

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

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

5.1 Detailed analysis

In this section, we design our L-level recommendation policy for $L > 3$. Similarly as Section 4, we first prove them in the case of 2 arms (Theorem 5.3 and Corollary 5.6). We then extend them to the case of constant number of arms (Theorem 5.7).

Now we start with the case of 2 arms. Our recommendation policy has L levels and two types of groups: G-groups and Γ-groups. Each level has σ^2 G-groups for $\sigma = 2^{10}\log(T)$. Label the G-groups in the l-th level as $G_{l,u,v}$ for $u,v \in [\sigma]$. Level 2 to level L also have σ^2 Γ-groups. Label the Γ-groups in the l-th level as $\Gamma_{l,u,v}$ for $u,v \in [\sigma]$. Each first-level group ($G_{1,u,v}$ for $u,v \in [\sigma]$) has T_1 full-disclosure path of L_K^{FDP} rounds in parallel. For $l \geq 2$, there are T_l agents in group $G_{l,u,v}$ and there are $T_l(\sigma-1)$ agents in group $\Gamma_{l,u,v}$. We will pick $T_1,\ldots,T_L$ in the proof of Theorem 5.3.

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

Theorem 5.3. *The L-level recommendation policy gets regret*

$$O\left(T^{2^{L-1}/(2^L-1)}\log^2(T)\right) \quad \textit{for} \quad L \leq \log(\ln(T)/\log(\sigma^4)).$$

In particular, if we pick $L = \log(\ln(T)/\log(\sigma^4))$, the regret is $O(T^{1/2}\,\mathrm{polylog}(T))$.

Proof. Wlog we assume $\mu_1 \geq \mu_2$ as the recommendation policy is symmetric to both arms. We will set T_l's later in the proof. Before that, we are only going to assume $T_l/T_{l-1} \geq \sigma^4$ for $l = 2,\ldots,L-1$ and $T_1 \geq \sigma^4$.

Similarly as the proof of Theorem 4.2, 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 3.2, 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,τ_1,τ_2} to be the event that the mean of τ_1-th to τ_2-th pulls in the tape is at most $\sqrt{\frac{3\log(T)}{\tau_2-\tau_1+1}}$ away from μ_a. By Chernoff bound,

  $$\Pr[W_2^{t,a,\tau_1,\tau_2}] \geq 1 - 2\exp(-6\log(T)) \geq 1 - 2/T^6.$$

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

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

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

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

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

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

  $$\Pr[W_3^{l,u,a,high} | H_l = h_l] \geq (1 - \Phi(1/2)) - \frac{5}{\sqrt{T_l}} > 1/4.$$

  Similarly we also have

  $$\Pr[W_3^{l,u,a,low} | H_l = h_l] > 1/4$$

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

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

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

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

  Since this holds for all h_l's, we have $\Pr[W_3^{l,a}] \geq 1 - 1/T^2$. Finally define $W_3 = \bigcap_{l,a} W_3^{l,a}$. By union bound, we have

  $$W_3 \geq 1 - 2L/T^2.$$

- **Anti-concentration of the empirical mean of arm a pulls in the first level:**

  For first-level groups $G_{1,u,1},...,G_{1,u,\sigma}$ 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 $\varepsilon_0 = 1$, $\varepsilon_1 = \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}}$ and $\varepsilon_l = 1/(4\sqrt{T_l \sigma})$ for $l \geq 2$.

Claim 5.4. *For any arm a and $2 \leq l \leq L$, if $\mu_1 - \mu_2 \leq \varepsilon_{l-1}$, then for any $u \in [\sigma]$, there are at least T_l pulls of arm a in groups $G_{l,u,1}, G_{l,u,2}, \ldots, G_{l,u,\sigma}$ and there are at least $T_l \sigma(\sigma - 1)$ pulls of arm a in the l-th level Γ-groups.*

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

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

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

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

$$\begin{aligned}\bar{\mu}_a^t - \bar{\mu}_{3-a}^t &> \mu_a - \mu_{3-a} + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}} \\ &\geq -\varepsilon_1 + \frac{1}{4\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{4\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}} \\ &\geq \frac{1}{8\sqrt{N_{K,a}^{\mathsf{FDP}} T_1 \sigma}} + \frac{1}{8\sqrt{N_{K,3-a}^{\mathsf{FDP}} T_1 \sigma}}.\end{aligned}$$

By Assumption 2.1, we have

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

The last inequality holds since C_{est} is a small enough constant defined in Assumption 2.1. Therefore we know agents in groups $G_{2,1,u_a},\dots,G_{2,\sigma,u_a}$ and $\Gamma_{2,1,u_a},\dots,\Gamma_{2,\sigma,u_a}$ all pull arm a.

Now we consider the case when $l > 2$ and assume the claim is true for smaller l's. For each arm a, W_3 implies that there exists u_a such that $W_3^{l-1,u_a,a,high}$ and $W_3^{l-1,u_a,3-a,low}$ happen. Recall $n^{l-1,u_a,a}$ is the number of arm a pulls in groups $G_{l-1,u_a,1},\dots,G_{l-1,u_a,\sigma}$. The induction hypothesis implies that $n^{l-1,u_a,a} \geq T_{l-1}$. $W_3^{l-1,u_a,a,high}$ together with $n^{l-1,u_a,a} \geq T_{l-1}$ implies that the empirical mean of arm a pulls in group $G_{l-1,u_a,1},\dots,G_{l-1,u_a,\sigma}$ is at least $\mu_a + 1/\sqrt{n^{l-1,u_a,a}}$. For any agent t in groups $G_{l,1,u_a},\dots,G_{l,\sigma,u_a}$ and $\Gamma_{l,1,u_a},\dots,\Gamma_{l,\sigma,u_a}$, it observes history of groups $G_{l-1,u_a,1},\dots,G_{l-1,u_a,\sigma}$ and all groups in levels below level $l-1$. Notice that the groups in the first $l-2$ levels have at most $(T_1 L_K^{\mathsf{FDP}} + T_2 + \cdots + T_{l-2})\sigma^3 \leq T_{l-1}/(12\log(T)) \leq n^{l-1,u_a,a}/(12\log(T))$ agents. By W_2, we have

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

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

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

Then we have

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

The last inequality holds because $n^{l-1,u_a,a}$ and $n^{l-1,u_a,3-a}$ are at most $T_{l-1}\sigma$. By Assumption 2.1, we have

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

The last inequality holds since C_{est} is a small enough constant defined in Assumption 2.1. Therefore agents in groups $G_{l,1,u_a},\ldots,G_{l,\sigma,u_a}$ and $\Gamma_{l,1,u_a},\ldots,\Gamma_{l,\sigma,u_a}$ all pull arm a. ☐

Claim 5.5. *For any $2 \leq l \leq L$, if $\varepsilon_{l-1}\sigma \leq \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma$, there are no pulls of arm 2 in groups with level $l,\ldots,L$.*

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

- $\varepsilon_{l-1}\sigma \leq \mu_1 - \mu_2 \leq \varepsilon_{l-2}$ for $l = 2$(i.e. $\varepsilon_1\sigma \leq \mu_1 - \mu_2 \leq \varepsilon_0$): 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

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

  By Assumption 2.1, we have

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

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

- $\varepsilon_{l-1}\sigma \le \mu_1 - \mu_2 \le \varepsilon_{l-2}$ for $l > 2$: By claim 5.4, for any agent t in level at least l, that agent will observe at least T_{l-1} arm a pulls. By W_2, we have

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

By Assumption 2.1, we have

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

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

- $\varepsilon_{l-2} < \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma$ for $l = 3$ (i.e. $\varepsilon_1 < \mu_1 - \mu_2 < \varepsilon_1\sigma$): By Claim 5.4, 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 2.1, we have

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

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

- $\varepsilon_{l-2} < \mu_1 - \mu_2 < \varepsilon_{l-2}\sigma$ for $l > 3$: Since $\mu_1 - \mu_2 < \varepsilon_{l-2}\sigma < \varepsilon_{l-3}$, by Claim 5.4, for any agent t in level at least l, that agent will observe at least $T_{l-2}\sigma^2$ arm a pulls (just from level $l-2$). By W_2, we have

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

By Assumption 2.1, we have

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

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

□

Now we set the group sizes T_l's as following. For $l < L$,

$$T_l = T^{\frac{2^{L-1}+2^{L-2}+\cdots+2^{L-l}}{2^{L-1}+2^{L-2}+\cdots+1}}/\sigma^3.$$

and

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

We restrict L to be at most $\log(\ln(T)/\log(\sigma^4))$ so that $T_l/T_{l-1} \geq T^{1/2^L} \geq \sigma^4$ for $l = 2, \ldots, L-1$. T_L is a little bit different because we want total number of agents to be T.

By Claim 5.5, we know that the regret conditioned the intersection of clean events is at most

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

□

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

Corollary 5.6. *With the proper setting of L and $T_1, \ldots, T_L$ described above, the L-level recommendation policy gets regret $O(\min(1/\Delta, T^{1/2})\,\mathrm{polylog}(T))$. Here $\Delta = |\mu_1 - \mu_2|$ and the L-level recommendation policy does not need to know Δ. Moreover, agent t observes a subhistory of size at least $\Omega(\lfloor t/\mathrm{polylog}(T)\rfloor)$.*

Proof. Notice that in the proof of Theorem 5.3, by the end of Claim 5.5, the only constraint we need about T_l's is that $T_l/T_{l-1} \geq \sigma^4$ for $l = 2, ..., L-1$ and $T_1 \geq \sigma^4$. And our new settings of T_l's still satisfy this constraint. So we can reuse the proof of Theorem 5.3 till the end of Claim 5.5.

Recall in the proof of Theorem 5.3, $\varepsilon_l = \Theta(1/\sqrt{T_l\sigma})$ for $l \in [L-1]$ and $\varepsilon_0 = 1$. Consider two cases:

- $\Delta < \varepsilon_{L-1}\sigma$. In this case, notice that even always picking the sub-optimal arm gives expected regret at most $T(\mu_1 - \mu_2) = T\Delta = O(T^{1/2}\text{polylog}(T))$. On the other hand, $T^{1/2} = O(\text{polylog}(T)/\Delta)$. Therefore, the regret is $O(\min(1/\Delta, T^{1/2})\,\text{polylog}(T))$.

- $\Delta \geq \varepsilon_{L-1}\sigma$. In this case, we can find $l \in \{2, ..., L\}$ such that $\varepsilon_{l-1}\sigma \leq \Delta < \varepsilon_{l-2}\sigma$. By Claim 5.5, we can upper bound the regret by

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

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

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

Here we discuss about how to extend Theorem 5.3 and Corollary 5.6 to the case when K is a constant larger than 2. As the proof is very similar to the proofs of Theorem 5.3 and Corollary 5.6, we only provide a proof sketch of what changes to make.

Theorem 5.7. *Theorem 5.3 and Corollary 5.6 can be extended to the case when K is constant larger than 2. In the extension of Corollary 5.6, Δ is defined as the difference between means of the best and the second best arm.*

Proof Sketch. We still wlog assume arm 1 has the highest mean (i.e. $\mu_1 \geq \mu_a, \forall a \in \mathcal{A}$. We first extend the clean events (i.e. W_1, W_2, W_3, W_4) in Theorem 5.3 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 $W_3^{l,a} = \bigcup_u \left(W_3^{l,u,a,high} \cap \left(\bigcap_{a'\neq a} W_3^{l,u,a',low}\right)\right)$ and $W_3 = \bigcap_{l,a} W_3^{l,a}$. We extend W_4 in a similar way: define W_4^a as $\bigcup_u \left(W_4^{u,a,high} \cap \left(\bigcap_{a'\neq a} W_4^{u,a',low}\right)\right)$ 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 5.4 and Claim 5.5. The statement of Claim 5.4 should be changed to "For any arm a and $2 \le l \le L$, if $\mu_1 - \mu_a \le \varepsilon_{l-1}$, then for any $u \in [\sigma]$, there are at least T_l pulls of arm a in groups $G_{l,u,1}, G_{l,u,2}, ..., G_{l,u,\sigma}$ and there are at least $T_l\sigma(\sigma - 1)$ pulls of arm a in the l-th level Γ-groups". The statement of Claim 5.5 should be changed to "For any $2 \le l \le L$, if $\varepsilon_{l-1}\sigma \le \mu_1 - \mu_a < \varepsilon_{l-2}\sigma$, there are no pulls of arm a in groups with level $l, ..., L$."

The proof of Claim 5.5 can be easily changed to prove the new version by changing "arm 2" to "arm a". The proof of Claim 5.4 needs some additional argument. In the proof of Claim 5.4, 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 $a' \neq a$. The proof of Claim 5.4 goes through if $\mu_1 - \mu_{a'} \le \varepsilon_{l-2}$ since then there will be enough arm a' pulls in level $l - 1$. We need some additional argument for the case when $\mu_1 - \mu_{a'} > \varepsilon_{l-2}$. Since $\mu_1 - \mu_{a'} > \varepsilon_{l-2} > \varepsilon_{l-1}\sigma$, we can use the same proof of Claim 5.5 (which rely on Claim 5.4 but for smaller l's) to show that there are no arm a' pulls in level l 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 proofs of Theorem 5.3 and Corollary 5.6 bound it by consider the regret from pulling the suboptimal arm (i.e. arm 2). When extending to more than 2 arms, we can do the exactly same argument for all arms except arm 1. This will blow up the regret by a factor of $(K - 1)$ which is a constant. □

Acknowledgment

We would like to thank Robert Kleinberg for discussions in the early stage of this project.

References

[1] 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.

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

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. Preliminary version in *36th IEEE FOCS*, 1995.

[4] 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*.

[5] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multiarmed bandit mechanisms. *SIAM J. on Computing (SICOMP)*, 43(1):194–230, 2014. Preliminary version in *10th ACM EC*, 2009.

[6] 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*.

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

[8] 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.

[9] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *CoRR arXiv:1704.09011*, 2017. Working paper (2017-2019). Major revision in *Management Science* (2019).

[10] Dirk Bergemann and Stephen Morris. Information design, bayesian persuasion, and bayes correlated equilibrium. *American Economic Review*, 106(5):586–591, 2016.

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

[12] 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.

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

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

[15] 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.

[16] Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *Quarterly Journal of Economics*, 133(2):871925, 2018. Working paper since 2013, titled 'Optimal design for social learning'.

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

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

[19] 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.

[20] 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.

[21] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[22] 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*.

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

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

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

[26] Sampath Kannan, Michael J. Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *ACM Conf. on Economics and Computation (ACM EC)*, pages 369–386, 2017.

[27] 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)*, pages 2231–2241, 2018.

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

[29] 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.

[30] 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.

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

[32] 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*.

[33] Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[34] T. Lattimore and Cs. Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

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

[36] 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.

[37] 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.

[38] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 2020. To appear; preliminary version in *ACM EC 2015*.

[39] 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*. To appear in *Operations Research* after a revision.

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

[41] Afshin Nikzad. Persuading a pessimist: Simplicity and robustness. 2019.

[42] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *Ann. Statist.*, 44(2):660–681, 04 2016.

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

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

[45] Mark Sellke and Aleksandrs Slivkins. Sample complexity of incentivized exploration., 2020. Working paper, available at arxiv.org/abs/2002.00558.

[46] 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.

[47] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. Also available at https://arxiv.org/abs/1904.07272.

[48] Ina Taneva. Information design, 2016. Working paper. Forthcoming in *American Economic Journal: Macroeconomics*.