Proceedings of Machine Learning Research vol 201:1–37, 2023 34th International Conference on Algorithmic Learning Theory

Max-Quantile Grouped Infinite-Arm Bandits

Ivan Lau IVAN.LAU@RICE.EDU
Rice University

Yan Hao Ling LINGYH@NUS.EDU.SG
National University of Singapore

Mayank Shrivastava MAYANKS4@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Jonathan Scarlett SCARLETT@COMP.NUS.EDU.SG
National University of Singapore

Editors: Shipra Agrawal and Francesco Orabona

Abstract

In this paper, we consider a bandit problem in which there are a number of groups each consisting of infinitely many arms. Whenever a new arm is requested from a given group, its mean reward is drawn from an unknown reservoir distribution (different for each group), and the uncertainty in the arm's mean reward can only be reduced via subsequent pulls of the arm. The goal is to identify the infinite-arm group whose reservoir distribution has the highest $(1-\alpha)$-quantile (e.g., median if $\alpha = \frac{1}{2}$), using as few total arm pulls as possible. We introduce a two-step algorithm that first requests a fixed number of arms from each group and then runs a finite-arm grouped max-quantile bandit algorithm. We characterize both the instance-dependent and worst-case regret, and provide a matching lower bound for the latter, while discussing various strengths, weaknesses, algorithmic improvements, and potential lower bounds associated with our instance-dependent upper bounds.

Keywords: Bandit algorithms, infinite-arm bandits, grouped bandits, lower bounds

1. Introduction

Multi-armed bandit (MAB) algorithms are widely adopted in scenarios of decision-making under uncertainty (Slivkins, 2019; Lattimore and Szepesvári, 2020). In theoretical MAB studies, two particularly common performance goals are *regret minimization* and *best arm identification*, and this paper is more related to the latter. The most basic form of best arm identification seeks to identify a *single* arm with the highest mean reward from a *finite* set of arms (Kaufmann et al., 2016). Other variations include but not limited to (i) identifying *multiple arms* with highest mean rewards (Kalyanakrishnan et al., 2012); (ii) identifying a single arm with a near-maximal mean reward from an *infinite* set of arms (Aziz et al., 2018; Chaudhuri and Kalyanakrishnan, 2018); (iii) identifying, for each group of arms, a single arm with the highest mean reward (Gabillon et al., 2011; Bubeck et al., 2013); (iv) identifying a single arm-group from a set of groups (of arms) whose worst arm has the highest mean reward (Wang and Scarlett, 2021).

In this paper, we consider a novel setup featuring aspects of both settings (ii) and (iv) above. Specifically, we are given a finite number of groups, with each group having an (uncountably) infinite number of arms, and our goal is to identify the group whose top $(1-\alpha)$-*quantile arm* (in terms of the mean reward) is as high as possible. While this setup builds on existing ones, we will find that it comes with unique challenges, with non-minor differences in both the upper and lower bounds.

A natural motivation for this problem is *comparing large populations* (e.g., of users, items, or otherwise). As a concrete example, in recommendation systems, one may wish to know which among several populations has the highest median click-through rate, while displaying as few total recommendations as possible. Beyond any specific applications, we believe that our problem is natural to consider in the broader context of infinite-arm bandits.

Before formally introducing the problem and stating our contributions, we outline some related work.

1.1. Related Work

The related work on multi-armed bandits is extensive (e.g., see (Slivkins, 2019; Lattimore and Szepesvári, 2020) and the references therein); we only provide a brief outline here, with an emphasis on only the most closely related works.

Finite-arm settings. The standard (single) best arm identification problem was studied in (Audibert et al., 2010; Gabillon et al., 2012; Jamieson and Nowak, 2014; Kaufmann et al., 2016; Garivier and Kaufmann, 2016), among others. These works are commonly distinguished according to whether the time horizon is fixed (fixed-budget setting) or the target error probability is fixed (fixed-confidence setting), and our focus is on the latter. In particular, we will utilize anytime confidence bounds from (Kaufmann et al., 2016) for upper-bounding the number of pulls.

A *grouped* best-arm identification problem was studied in (Gabillon et al., 2011; Bubeck et al., 2013; Scarlett et al., 2019), where the arms are allocated into groups, and the goal is to find the best arm in each group. Another notable setting in which multiple arms are returned is that of subset selection, where one seeks to find a subset of k arms attaining the highest mean rewards (Kalyanakrishnan et al., 2012; Kaufmann and Kalyanakrishnan, 2013; Kaufmann et al., 2016). Group structure can also be incorporated into *structured bandit* frameworks (Huang et al., 2017; Gupta et al., 2020; Mukherjee et al., 2020; Neopane et al., 2021). Perhaps the most related among these is that of max-min grouped bandits (Wang and Scarlett, 2021), which seeks a group whose worst arm is as high as possible in a finite-arm setting. We discuss the main similarities and differences to our setting in Appendix F, as well as highlighting a connection to structured best-arm identification (Huang et al., 2017).

Infinite-arm settings. One line of works in infinite-arm bandits assumes that the actions lie in a metric space, and the associated mean rewards satisfy some smoothness condition such as being locally Lipschitz (Kleinberg et al., 2008; Bubeck et al., 2008, 2011; Grill et al., 2015; Kleinberg et al., 2019). More relevant to our paper is the line of works considering a *reservoir distribution* on arms; throughout the learning process, new arms can be requested and previously-requested ones can be pulled. Earlier works in this direction focused primarily on regret minimization (e.g., (Berry et al., 1997; Wang et al., 2008; Bonald and Proutière, 2013; David and Shimkin, 2014; Li and Xia, 2017; Kalvit and Zeevi, 2021)), and several recent works have considered pure exploration (e.g., (Carpentier and Valko, 2015; Jamieson et al., 2016; Chaudhuri and Kalyanakrishnan, 2018; Aziz et al., 2018; Chaudhuri and Kalyanakrishnan, 2019; Ren et al., 2019; Katz-Samuels and Jamieson, 2020; Zhang and Ong, 2021)). Throughout the paper, we particularly focus on the work (Aziz et al., 2018), which studies the problem of finding a single arm in the top $(1-\alpha)$-quantile. See Appendix F for a discussion of some key similarities and differences.

The idea of measuring the performance against a quantile has appeared in several of these works, including (Chaudhuri and Kalyanakrishnan, 2018; Aziz et al., 2018). Related notions include finding a "good" arm in a scenario where there are two types of arm (Jamieson et al., 2016), and finding a subset of a top fraction of arms in a finite-arm setting (Chaudhuri and Kalyanakrishnan, 2019; Ren et al., 2019). The notion of finding an optimal quantile has also appeared extensively in risk-aware bandits (Tan et al., 2022), but these problems consist of a single set of arms and consider quantiles of the reward distributions, which is handled very differently to quantiles of reservoir distributions.

2. Problem Setup and Contributions

Arms and rewards. We first describe the problem aspects that are the same as regular MAB problems. We are given a collection $\mathcal{A}$ of arms, which has some unknown mean-reward mapping $\mu\colon \mathcal{A} \to \mathbb{R}$. In each round, indexed by $t \geqslant 1$, the algorithm pulls one or more arms[1] in $\mathcal{A}$ and observes their corresponding rewards. We consider the stochastic reward setting, in which for each arm $j \in \mathcal{A}$, the observations of its reward $\{X_{j,t}\}_{t\geqslant 1}$ are i.i.d. random variables from some distribution with mean $\mu_j := \mu(j)$. It is also useful to define the empirical mean of arm j at round t, defined as

$$\hat{\mu}_{j,T_j(t)} := \frac{1}{T_j(t)} \sum_{\substack{\tau \in \{1,\dots,t\}: \\ \text{arm } j \text{ pulled}}} X_{j,\tau}, \tag{1}$$

where $T_j(t) \leqslant t$ is the number of pulls of arm j up to round t. In standard best-arm identification problems, the goal is to identify the best arm with high probability using as few pulls as possible.

We will restrict our attention to classes of arm distributions that are *uniquely parametrized by their mean*, e.g., Bernoulli(μ) or $N(\mu, \sigma^2)$ with $\sigma^2 > 0$ being the same for every arm. By doing so, the notion of a reservoir distribution (see below) can be introduced using probability distributions on $\mathbb{R}$, which is significantly more convenient compared to more general distributions.

Infinite-arm and group notions. In our setup, the number of arms $\mathcal{A}$ is (potentially uncountably) infinite. Furthermore, these arms are partitioned into a finite number of *disjoint* groups, with each group having a (potentially uncountably) infinite number of arms. The set of these disjoint groups is denoted by $\mathcal{G}$. The mean rewards for each group $G \in \mathcal{G}$ form a probability space $(M_G, \mathcal{F}_G, P_G)$, where $M_G = \mu(G) = \{\mu_j\colon j \in G\}$. For each group G, we define a corresponding *reservoir distribution* CDF $F_G\colon M_G \to [0,1]$, as well as a quantile function $F_G^{-1}\colon [0,1] \to M_G$, by

$$F_G(\tau) := P_G(\{\mu \leqslant \tau\}) \quad \text{and} \quad F_G^{-1}(p) := \inf\{\mu : F_G(\mu) \geqslant p\}, \tag{2}$$

1. We will find it useful to let t index "rounds" each possibly consisting of several arm pulls, but we will still be interested in the *total* number of arm pulls.

where F_G is assumed to have a bounded support. Our goal is to identify a group in $\mathcal{G}$ with the highest $(1-\alpha)$-quantile, i.e., a group G satisfying

$$F_G^{-1}(1-\alpha) = \max_{G' \in \mathcal{G}} F_{G'}^{-1}(1-\alpha). \tag{3}$$

Observe that when α is close to zero, this problem resembles that of finding a near-maximal arm across all groups, whereas when α is close to one, the problem resembles the max-min problem of finding the group whose worst arm is as high as possible (Wang and Scarlett, 2021). Throughout the paper, we treat $\alpha \in (0,1)$ as a fixed constant (e.g., $\alpha = \frac{1}{2}$ for the median), meaning its dependence may be omitted in $O(\cdot)$ notation.

Throughout the course of the algorithm, the following can be performed:

- The algorithm can request to receive one or more additional arms from any group of its choice; if that group is $G \in \mathcal{G}$, then the arm mean is drawn according to F_G.
- Among all the arms requested so far, the algorithm can perform *pulls* of the arms and observe the corresponding rewards, as usual.

We are interested in keeping the total number of arm pulls low; the total number of arms requested is not of direct importance (similar to previous infinite-arm problems such as (Aziz et al., 2018)).

It will be convenient to uniquely index each arm in a given group G by $j \in [0,1]$ such that any new arm drawn from the reservoir distribution has its j value drawn uniformly from $[0,1]$, and has mean reward $\mu_{G,j} = F_G^{-1}(j)$.[2] We will use this convention, but it is important to note that whenever the algorithm requests an arm, its index j remains unknown. We also note that two different indices $j \neq j'$ in a given group could still have the same means (i.e., $\mu_{G,j} = \mu_{G,j'}$), e.g., when the underlying reservoir distribution is discrete.

ϵ, Δ-relaxations. When no assumptions are made on the reservoir distributions, one may encounter scenarios where the $(1-\alpha)$-quantile of a given group is arbitrarily hard to pinpoint (e.g., because the underlying CDF has a near-horizontal region, implying a very low probability of any given arm being near the precise quantile). To alleviate this challenge, we add an ϵ-relaxation on α for some $\epsilon < \min(\alpha, 1-\alpha)$ to limit the effort spent on identifying a quantile that is hard to pinpoint. In particular, we relax the task into finding a group G satisfying $F_G^{-1}(1-\alpha+\epsilon) \geqslant \max_{G' \in \mathcal{G}} F_{G'}^{-1}(1-\alpha-\epsilon)$. Moreover, we further relax the task by only requiring the chosen group G satisfies

$$F_G^{-1}(1-\alpha+\epsilon) \geqslant \max_{G' \in \mathcal{G}} F_{G'}^{-1}(1-\alpha-\epsilon) - \Delta, \tag{4}$$

for some $\Delta > 0$. This relaxation allows us to limit the effort on distinguishing arms (potentially from different groups) whose expected rewards are very close to each other; analogous relaxations are common in standard best-arm identification problems. The general goal of the algorithm for our setup is to identify a group satisfying (4) with high probability while using as few arm pulls as possible.

Summary of contributions With the problem setup now in place, we can now summarize our main contributions (beyond the problem formulation itself):

2. We may assume that the same j value is never drawn twice, since this is a zero-probability event. We emphasize that with $\mu = F_G^{-1}(j)$, the assumption of unique j values does not necessarily imply unique arm means, as we still allow F_G to have mass points.

- We introduce a two-step algorithm based on first requesting a fixed number of arms for each group, and then running a finite-arm subroutine. Our main result for this algorithm is an instance-dependent upper bound on the number of pulls to guarantee (4) with high probability (Corollary 4), expressed in terms of the reservoir distributions and fundamental gap quantities introduced in Section 3.3.
- In addition, we establish a worst-case upper bound in terms of ϵ and Δ alone (Eq. (13)), and a guarantee for the finite-arm setting (Theorem 3). We believe that these should be of independent interest.
- We show that adapting our two-step algorithm to a multi-step algorithm can lead to a better instance-dependent bound (Corollary 5), and discuss how the resulting gap terms may be similar to those of a potential instance-dependent lower bound (Appendix F.3), though formalizing the latter is left for future work.
- By deriving a worst-case lower bound (Theorem 6) and comparing it with our upper bound, we show that for worst-case instances, the optimal number of arm pulls scales as $\frac{|\mathcal{G}|}{\Delta^2\epsilon^2}$ (for $|\mathcal{G}| \geqslant 2$) up to logarithmic factors.

Some of the innovations in our analysis include (i) suitably identifying the relevant "gaps" in the finite-arm elimination algorithm and adapting the analysis accordingly; (ii) setting up the relaxation (4) and determining how this impacts the number of arms to request and how to characterize the high-probability behavior of their gaps; (iii) suitably extending the two-step algorithm to a multi-step algorithm; and (iv) proving the lower bound via a likelihood-ratio based analysis that is distinct from others in the bandit literature to the best of our knowledge.

3. Algorithm and Upper Bound

In this section, we introduce our main algorithm and provide its performance guarantee. We make the following standard assumptions on the reward distributions.

Assumption 1 *For every arm j (in every group G; the group dependence is left implicit here), we assume that the mean reward μ_j is bounded in $[0,1]$,[3] and that the reward distribution is sub-Gaussian with parameter $\sigma^2 \leqslant 1$.[4] That is, for each arm j, and for each $\lambda \in \mathbb{R}$, if X is a random variable drawn from the arm's reward distribution, then $\mathbb{E}[X] = \mu_j$ and $\mathbb{E}[e^{\lambda(X-\mu_j)}] \leqslant \exp(\lambda^2\sigma^2/2)$. Moreover, we assume that all of the reward distributions come from a common family of distributions that are uniquely parametrized by their mean (e.g., Bernoulli, or Gaussian with a fixed variance).*

3.1. Description of the Algorithm

We present our two-step algorithm in Algorithm 1. This algorithm uses a finite-arm best-quantile identification (BQID) algorithm `FiniteArmBQID` as a sub-routine, and its description is deferred to Algorithm 2 in Appendix C.3. For now, we only need to treat this step in a "black-box" manner with certain guarantees outlined at the end of this subsection.

3. Any finite interval can be shifted and scaled to this range.
4. The upper bound $\sigma^2 \leqslant 1$ is for convenience in applying known confidence bounds, and can easily be relaxed.

Our algorithm takes a two-step approach of first requesting a fixed number of arms from each group, and then running a finite-arm algorithm. This approach was taken in (Aziz et al., 2018) for the (non-grouped) problem of finding an arm within a given quantile, but the results and analysis turn out to be quite different; see Appendix F.3 for further discussion. In Section 3.5, we will discuss further improvements via a multi-step approach.

Algorithm 1 Main Algorithm

Require: Finite set of infinite arm groups $\mathcal{G}$, parameters $\alpha, \epsilon, \Delta, \delta \in (0,1)$ where $\delta < \epsilon < \min(\alpha, 1-\alpha)$

1: $N := N(\epsilon, \delta) = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil$
2: **for** each $G \in \mathcal{G}$ **do**
3:     Request N arms from G according to F_G to form a (random) finite arm group H
4: Let $\mathcal{H} = \left\{H_1, H_2, \ldots, H_{|\mathcal{G}|}\right\}$ be a finite set of finite arm groups
5: Run `FiniteArmBQID` (Algorithm 2 in Appendix C.3) with input $(\mathcal{H}, \alpha, \Delta, \delta)$ to identify a group $\widehat{H} \in \mathcal{H}$
6: Return the group $\widehat{G}$ corresponding to $\widehat{H}$.

First, for each group $G \in \mathcal{G}$, we randomly and independently request $N = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil$ arms from the reservoir distribution F_G to form a (random) arm group H of size N. This yields a set $\mathcal{H} = \left\{H_1, H_2, \ldots, H_{|\mathcal{G}|}\right\}$ of finite-arm groups, each with means denoted by $\{\mu_{H,j}\}_{j \in H}$ with continuous indexing $j \in [0,1]$ as discussed in Section 2. We make use of the corresponding (discrete) probability spaces $(M_H, \mathcal{F}_H, P_H)$ of mean rewards, with CDFs and quantile functions given by

$$F_H(\tau) := P_H(\{\mu \leqslant \tau\}) = \frac{1}{|H|} \sum_{j \in H} \mathbf{1}\{\mu_{H,j} \leqslant \tau\} \quad \text{and} \quad F_H^{-1}(p) := \inf\{\mu : F_H(\mu) \geqslant p\}. \tag{5}$$

In Line 5 of Algorithm 1, our requirement on the finite-arm subroutine `FiniteArmBQID` is that it identifies a group $\widehat{H} \in \mathcal{H}$ with a $(1-\alpha)$-quantile that is at most Δ-suboptimal:

$$F_{\widehat{H}}^{-1}(1-\alpha) \geqslant \max_{H \in \mathcal{H}} F_H^{-1}(1-\alpha) - \Delta. \tag{6}$$

We then simply return the infinite-arm group $\widehat{G}$ corresponding to $\widehat{H}$ (the two have a trivial one-to-one mapping). The high-level ideas behind the (elimination-based) algorithm we use for `FiniteArmBQID` will be introduced in Section 3.3, and the full details will be given in Appendix C.

3.2. Correctness

Our theoretical analysis is done via a series of intermediate results, and we defer the proof details to Appendix A. We will frequently make use of two high-probability events, defined as follows:

- Event A: For each $G \in \mathcal{G}$ and the corresponding finite group H (generated in Line 3 of Algorithm 1), we have

$$F_G^{-1}(1-\alpha-\epsilon) \leqslant F_H^{-1}(1-\alpha) \leqslant F_G^{-1}(1-\alpha+\epsilon). \tag{7}$$

  That is, the mean reward of the $(1-\alpha)$-quantiles of the sampled arms H is between the mean rewards of the $(1-\alpha \pm \epsilon)$-quantile of the arms in G.

- Event B: The group $\widehat{H}$ returned by Algorithm 2, i.e., Line 5 of Algorithm 1 satisfies (6).

In Appendix A, we show that these each hold with probability at least $1 - \delta$, and then deduce the following correctness guarantee.

Theorem 2 *Consider Algorithm 1 with inputs $(\mathcal{G}, \alpha, \epsilon, \Delta, \delta)$ as defined in Section 2. Under Assumption 1, with probability at least $1 - 2\delta$, Algorithm 1 identifies a group $G \in \mathcal{G}$ satisfying* (4).

Here and subsequently, the fact that `FiniteArmBQID` is performed via Algorithm 2 (Appendix C.3) is left implicit in most of our formal statements.

3.3. Number of Arm Pulls in Finite-Arm Subroutine

The arm pulls in Algorithm 1 occur entirely within the `FiniteArmBQID` subroutine (Algorithm 2 in Appendix C.3). This algorithm and its analysis will be generally similar to (Wang and Scarlett, 2021, Algorithm 1), though with different details. Their algorithm can be viewed as taking $\alpha = 1$ and $\Delta = 0$ in our notation, whereas we are interested in an *arbitrary* $\alpha \in (0, 1)$ and we allow for Δ to be positive. Moreover, in (Wang and Scarlett, 2021) the groups are allowed to overlap; we avoided such considerations because they seem tricky to formulate neatly in the infinite-arm setting, but in the finite-arm setting of this section it would be straightforward to incorporate. However, we emphasize that the key distinction is not these details, but rather the challenges of incorporating the finite-arm algorithm into an infinite-arm framework.

A complete description of the finite-arm subroutine is given in Algorithm 2 in Appendix C, and its analysis is given in Appendix D. While the details are deferred to later, we are in a position to proceed here with a "black-box" statement of its number of arm pulls, along with some intuition on why certain instance-dependent quantities arise. Algorithm 2 takes a finite set $\mathcal{H}$ of finite arm-groups and parameters $\alpha, \Delta, \delta \in (0, 1)$ as input, and seeks to output a group $H \in \mathcal{H}$ satisfying (6) with probability at least $1 - \delta$. It is based on successive elimination[5], and as with previous algorithms of this kind, it stops pulling a given arm altogether once it is "no longer of interest". To formalize this, each arm $j \in H$ is assigned a gap value $\Delta_{H,j}$, and we will show that once the empirical means and true means are at most $\Delta_{H,j}/4$ apart, arm $j \in H$ can stop being pulled. The intuition behind why arms can be eliminated is as follows:

- It may occur that $\mu_{H,j}$ is already known accurately enough that any further accuracy provides no additional information about the group's $(1 - \alpha)$-quantile.
- It may be that H is known to be a suboptimal group, so *all* its arms can stop being pulled.
- It may be that the optimal group is found and the algorithm can terminate.
- It may be that all remaining groups are known to satisfy (6) and the algorithm can terminate.

These four cases will be associated with gaps denoted by $\Delta'_{H,j}$, Δ_H, Δ_0, and Δ respectively (to be defined shortly), leading to the following overall gap for arm $j \in H$:

$$\Delta_{H,j} := \max\left\{\Delta, \Delta_H, \Delta_0, \Delta'_{H,j}\right\} \in [\Delta, 1]. \tag{8}$$

5. We expect that other approaches such as LUCB-type algorithms could also provide similar guarantees. However, see Appendix F.2 for a discussion on how an existing LUCB-based bound for a general structured bandit framework can be weaker than ours.

We now proceed formally. Let H^* be a group with a highest $(1-\alpha)$-quantile arm,[6] i.e., $H^* \in \operatorname{argmax}_{H \in \mathcal{H}} F_H^{-1}(1-\alpha)$. If multiple groups satisfy this condition, then H^* denotes an arbitrary single one of them. For each group H, we define the reward-gap Δ_H of group H as the difference between the $(1-\alpha)$-quantile arm of H^* and the $(1-\alpha)$-quantile arm of group H:

$$\Delta_H := F_{H^*}^{-1}(1-\alpha) - F_H^{-1}(1-\alpha). \tag{9}$$

Similarly, Δ_0 indicates the difference between the $(1-\alpha)$-quantile arms of group H^* and the remaining groups $H \in (\mathcal{H} \backslash H^*)$, i.e.,

$$\Delta_0 := \min_{H \in \mathcal{H}, H \neq H^*} \Delta_H = F_{H^*}^{-1}(1-\alpha) - \max_{H \in \mathcal{H}, H \neq H^*} F_H^{-1}(1-\alpha). \tag{10}$$

Note that $\Delta_0 > 0$ if and only if there is a unique optimal group. Finally, $\Delta'_{H,j}$ is the difference between the mean reward of arm j and the $(1-\alpha)$-quantile arm of the group it is in, i.e.,

$$\Delta'_{H,j} := \left|\mu_{H,j} - F_H^{-1}(1-\alpha)\right|. \tag{11}$$

The remaining term Δ in (8) is already specified as part of the problem, and is included to relax the success criterion for very challenging instances where $\max\{\Delta_H, \Delta_0, \Delta'_{H,j}\} \approx 0$ or even $\max\{\Delta_H, \Delta_0, \Delta'_{H,j}\} = 0$. Having defined $\Delta_{H,j}$, we now state an upper bound on the total number of arm pulls by `FiniteArmBQID`.

Theorem 3 *Let $(\mathcal{G}, \alpha, \epsilon, \Delta, \delta)$ be a valid input of Algorithm 1 and $\mathcal{H}$ be the set of random groups formed in Lines 2-3 of Algorithm 1. There exists a choice of* `FiniteArmBQID` *(see Algorithm 2 in Appendix C.3) yielding the following: Under Assumption 1, with probability at least $1-\delta$, the algorithm identifies a group $H \in \mathcal{H}$ satisfying (6) and uses a total number of arm pulls upper bounded by*

$$T(\epsilon, \delta, \Delta) \leqslant \sum_{H \in \mathcal{H}} \sum_{j=1}^{N} \frac{c}{\Delta_{H,j}^2} \log\left(\frac{|\mathcal{H}|N}{\delta} \log \frac{1}{\Delta_{H,j}^2}\right) \tag{12}$$

$$\leqslant d\left(\frac{|\mathcal{G}|}{\epsilon^2\Delta^2}\right)\left(\log^2\frac{|\mathcal{G}|}{\delta} + \left(\log\frac{|\mathcal{G}|}{\delta}\right)\left(\log\log\frac{1}{\Delta}\right)\right) \tag{13}$$

where c and d are universal constants, the gaps $\Delta_{H,j}$ as defined in (8) are random variables, and $N = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil$ is as given in Line 1 of Algorithm 1.

Proof The first line follows from Theorem 13 in Appendix C.3, whose proof is given in Appendix D. The second line lower bounds each gap by Δ, and performs asymptotic simplifications along with $|\mathcal{G}| = |\mathcal{H}|$ and the assumption $\delta < \epsilon$ specified in Algorithm 1. ■

6. Due to the randomness in generating the finite-arm groups, H^* does not necessarily coincide with the optimal infinite-arm group.

3.4. Number of Arm Pulls Without $\mathcal{H}$ Dependence

Since each finite group $H \in \mathcal{H}$ of arms consists of arms randomly sampled from their corresponding G, the gaps $\Delta_H, \Delta_0, \Delta'_{H,j}$ are random variables, and so $\Delta_{H,j} = \max\{\Delta, \Delta_H, \Delta_0, \Delta'_{H,j}\}$ are also random variables. Our next step is to bound these gaps with high probability.

For Δ_H and Δ_0, Event A above readily yields the following lower bounds:

$$\Delta_H \geqslant \widetilde{\Delta}_G := \max_{G' \in \mathcal{G}} F_{G'}^{-1}(1-\alpha-\epsilon) - F_G^{-1}(1-\alpha+\epsilon) \tag{14}$$

$$\Delta_0 \geqslant \widetilde{\Delta}_0 := \max_{G \in \mathcal{G}} F_G^{-1}(1-\alpha-\epsilon) - \max_{G' \in \mathcal{G},\, G' \neq G_\epsilon^*} F_{G'}^{-1}(1-\alpha+\epsilon), \tag{15}$$

where $G_\epsilon^* \in \mathrm{argmax}_{G \in \mathcal{G}} F_G^{-1}(1-\alpha+\epsilon)$. See Appendix B.1 for the details.

In contrast, based on the high-probability events established so far, there is no non-trivial lower bound on $\Delta'_{H,j}$. For example, if the mean rewards of all arms sampled for group H are concentrated around the $(1-\alpha)$-quantile, then $\Delta'_{H,j}$ will be small for each j. We address this difficulty by introducing further high-probability events under which $\Delta'_{H,j}$ can be suitably lower bounded.

Let $\mathcal{H} = \{H\}$ be the set of groups formed in Lines 2-3 of Algorithm 1. We partition the arms in H as follows. Let m be the smallest integer such that $(1-\alpha) - \left\lfloor \frac{1-\alpha}{\epsilon} \right\rfloor \epsilon + m\epsilon \geqslant 1$, i.e.,

$$m := \min\left\{k \in \mathbb{N} : k \geqslant \frac{\alpha}{\epsilon} + \left\lfloor \frac{1-\alpha}{\epsilon} \right\rfloor \right\} \in \left\{ \left\lfloor \frac{1}{\epsilon} \right\rfloor, \left\lceil \frac{1}{\epsilon} \right\rceil \right\} \geqslant 3, \tag{16}$$

where the lower bound of 3 follows since $\epsilon < \min(\alpha, 1-\alpha) \leqslant \frac{1}{2}$ (Line 1 of Algorithm 1). We partition H into $m+1$ disjoint subsets:

$$S_{H,i} := \begin{cases} H \cap [0, b_1) = \{j \in H \mid 0 \leqslant j < b_1\} & \text{for } i = 0, \\ H \cap [b_i, b_{i+1}) = \{j \in H \mid b_i \leqslant j < b_{i+1} = b_i + \epsilon\} & \text{for } 1 \leqslant i \leqslant m-1, \\ H \cap [b_m, 1] = \{j \in H \mid b_m \leqslant j \leqslant 1\} & \text{for } i = m, \end{cases} \tag{17}$$

where

$$b_i := (1-\alpha) - \left\lfloor \frac{1-\alpha}{\epsilon} \right\rfloor \epsilon + (i-1)\epsilon \quad \text{for } i = 1, 2, \dots, m. \tag{18}$$

That is, $[0, b_1), \dots, [1-\alpha-\epsilon, 1-\alpha), [1-\alpha, 1-\alpha+\epsilon), \dots, [b_m, b_{m+1} = 1]$ are $m+1$ disjoint intervals of $[0,1]$, and each interval has a size of at most ϵ. See Figure 1 for an illustration (in this example, $S_{H,0}$ is empty because $\frac{\alpha}{\epsilon}$ is an integer).

The idea in our subsequent analysis (leading to Corollary 4 below) is to bound the number of arms in each partition, and then bound the arm means in each partition by their "worst-case" values (i.e., those giving the smallest gaps). The partition width ϵ is chosen such that the resulting "rounding error" coincides with the desired guarantee (4) with parameter ϵ.

Specifically, in Appendix B.1, we use the above partitioning to show that $\Delta'_{H,j} \geqslant \widetilde{\Delta}'_{G,i}$, where

$$\widetilde{\Delta}'_{G,i} := \begin{cases} F_G^{-1}(1-\alpha-\epsilon) - F_G^{-1}(b_{i+1}) & \text{if } 0 \leqslant i < \left\lfloor \frac{1-\alpha}{\epsilon} \right\rfloor - 1 \\ F_G^{-1}(b_i) - F_G^{-1}(1-\alpha+\epsilon) & \text{if } \left\lfloor \frac{1-\alpha}{\epsilon} \right\rfloor + 1 < i \leqslant m \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

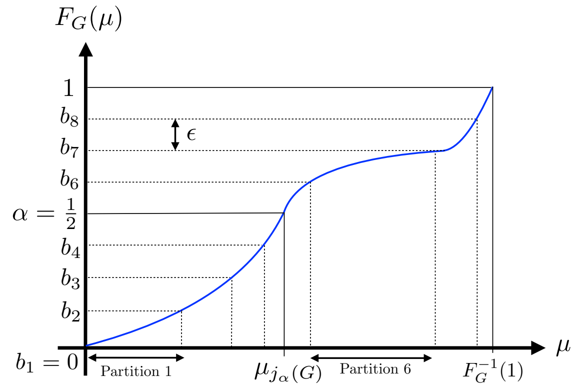

Figure 1: Example of partitioning of arms, with $\alpha = \frac{1}{2}$ and $\epsilon = \frac{1}{8}$.

Intuitively, the third case corresponds to arms so close to the quantile that no positive $\Delta'_{H,j}$ gap can be guaranteed (though other gaps such as Δ_H may still be positive).

Combining the lower bounds on Δ_H, Δ_0, and $\Delta'_{H,j}$, we have $\Delta_{H,j} \geqslant \max\{\Delta, \widetilde{\Delta}_G, \widetilde{\Delta}_0, \widetilde{\Delta}'_{G,i}\}$, where i is the index of the partition of H that j belongs to, i.e., $j \in S_{H,i}$. To simplify notation, let

$$\widetilde{\Delta}_{G,i} := \widetilde{\Delta}_{G(H),i} = \max\left\{\Delta, \widetilde{\Delta}_G, \widetilde{\Delta}_0, \widetilde{\Delta}'_{G,i}\right\} \leqslant \Delta_{H,j}. \tag{20}$$

To completely remove the dependence on $\mathcal{H}$, we also need to bound the number of arms in each partition $S_{H,i}$. In Lemma 10 in Appendix B.2, we use a concentration argument to establish that $|S_{H,i}| \leqslant 3\epsilon N$ for all i with probability at least $1 - \delta$, leading to our final guarantee stated as follows.

Corollary 4 *Let $(\mathcal{G}, \alpha, \epsilon, \Delta, \delta)$ be a valid input of Algorithm 1. Under Assumption 1, with probability at least $1 - 3\delta$, Algorithm 1 identifies a group $G \in \mathcal{G}$ satisfying (4), and uses a total number of arm pulls satisfying*

$$T(\epsilon, \delta, \Delta) \leqslant \sum_{G \in \mathcal{G}} \sum_{i=1}^{m} \frac{c}{\widetilde{\Delta}^2_{G,i}} \log\left(\frac{|\mathcal{G}|N}{\delta} \log \frac{1}{\widetilde{\Delta}^2_{G,i}}\right) \cdot 3\epsilon N \tag{21}$$

$$\leqslant d\left(\frac{|\mathcal{G}|}{\epsilon^2 \Delta^2}\right)\left(\log^2 \frac{|\mathcal{G}|}{\delta} + \left(\log \frac{|\mathcal{G}|}{\delta}\right)\left(\log\log \frac{1}{\Delta}\right)\right) \tag{22}$$

where c and d are universal constants, m is given in (16), $\widetilde{\Delta}_{G,i}$ is as defined in (20), and $N = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil$ as stated in Line 1 of Algorithm 1.

The proof is deferred to Appendix B.2. Here (21) is our main instance-dependent bound, whereas (22) is a weaker result obtained by lower bounding all gaps by Δ, as was done in (13) for the finite-arm setting (note that in both cases, the asymptotic simplifications rely on the assumption $\delta < \epsilon$ specified in Algorithm 1). In Section 4, we provide a near-matching algorithm-independent lower bound to (22) in the worst case, and discuss the difficulties in obtaining instance-dependent lower bounds that nearly match (21). Moreover, in Appendix F, we discuss some key differences compared to the problem of finding an arm in the top $(1 - \alpha)$-quantile of a single group (Aziz et al., 2018).

3.5. Further Improvement via a Multi-Step Algorithm

A potential weakness of the upper bound in (4) is that it is based on immediately choosing N large enough such that each group's quantile is ϵ-close to the true quantile in the sense of (7) (i.e., Event A). However, if a group is highly suboptimal, it could potentially be identified as such even if ϵ is replaced by a much higher value in the choice of N in Algorithm 1. Thus, if ϵ is very small, then immediately using it to choose N (which has strong $\frac{1}{\epsilon^2}$ dependence) may be highly wasteful.

To alleviate this problem, one may use a *sequence* of decreasing ϵ and Δ values. We denote these by $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_K)$ and $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_K)$ for some $K > 0$, and we assume that $(\epsilon_K, \Delta_K) = (\epsilon, \Delta)$. We proceed in K epochs, indexing them by k (initially $k = 1$) and proceeding as follows:

(i) Set $N_k = \big\lceil \frac{1}{2\epsilon_k^2} \log \frac{2|\mathcal{G}|}{\delta} \big\rceil$, and request N_k arms from each group.
(ii) Run successive elimination (`FiniteArmBQID`) with the finite sets of arms from Step (i) and parameter Δ_k.
(iii) If multiple groups remain and $k < K$, then increment k and return to Step (i) while permanently removing all groups that were eliminated in Step (ii) (as well as discarding all previously-requested arms).
(iv) Otherwise, return $\widehat{G}$ corresponding to the $\widehat{H}$ returned in the last invocation of `FiniteArmBQID`.

In the following, we think of K as having mild dependence with respect to the other parameters, e.g., if halving is used for the smaller of Δ and ϵ, then $K = O\big(\log \frac{1}{\min\{\epsilon,\Delta\}}\big)$.

We proceed by outlining how the analysis and results are affected. Since we are essentially running Algorithm 1 K times, the error probability increases from $O(\delta)$ to $O(\delta K)$. Since the dependence on δ is logarithmic, this effect of scaling δ is minimal. More importantly, we need to identify when suboptimal groups are eliminated. We let $\widetilde{\Delta}_G^{(\epsilon)}$ be defined as in (14) with an explicit dependence on ϵ, and introduce the following critical values:

- If there exists k such that $\widetilde{\Delta}_{G,i}^{(\epsilon_k)} > \Delta_k$, we let $k_{\max}(G)$ be the smallest such k value.
- If no such k exists, we let $k_{\max}(G) = K$.

Our analysis of Algorithm 1 reveals that a suboptimal group is eliminated when $\widetilde{\Delta}_G > \Delta$, and hence, the above definitions ensure that group G will be eliminated (with high probability) after the end of epoch $k_{\max}(G)$. As a result, we have the following generalization of Corollary 9.

Corollary 5 *Consider running the preceding multi-step algorithm with sequences $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_K)$ and $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_K)$ for some $K > 0$, with $\delta_k < \epsilon_k < \min(\alpha, 1-\alpha)$ for all k. Under Assumption 1, with probability at least $1 - 3K\delta$, the algorithm identifies a group $G \in \mathcal{G}$ satisfying (4), and uses a total number of arm pulls satisfying*

$$T(\epsilon, \delta, \Delta) \leqslant \sum_{G \in \mathcal{G}} \sum_{k=1}^{k_{\max}(G)} \sum_{i=1}^{m_k} \frac{c}{(\widetilde{\Delta}_{G,i}^{(\epsilon_k)})^2} \log\left(\frac{|\mathcal{G}| N_k}{\delta} \log \frac{1}{(\widetilde{\Delta}_{G,i}^{(\epsilon_k)})^2} \right) \cdot 3\epsilon_k N_k \tag{23}$$

where c is a universal constant, m_k is as in (16) with ϵ_k in place of ϵ, $\widetilde{\Delta}_{G,i}^{(\epsilon)}$ is as in (20) with $\Delta_G = \Delta_G^{(\epsilon)}$ explicitly depending on ϵ, and $N_k = \big\lceil \frac{1}{2\epsilon_k^2} \log \frac{2|\mathcal{G}|}{\delta} \big\rceil$.

This bound is more complicated than Corollary 4, but can give significant improvements due to highly suboptimal groups being eliminated early and avoiding the smallest ϵ_k terms. For example, in a scenario where all arms have the same gap term Δ but $\epsilon_{k_{\max}(G)} \gg \epsilon$ for all suboptimal G, the $\frac{1}{\epsilon^2}$ dependence in (22) would be improved to $\frac{1}{(\min_{G \neq G^*} \epsilon_{k_{\max}(G)})^2}$. See Appendix F.3 for further discussion, and Figure 2 (in Appendix G) for a visual illustration.

On the negative side, choosing the sequences $\boldsymbol{\epsilon}$ and $\boldsymbol{\Delta}$ may be tricky. For example, if we choose $\Delta_k = \Delta$ for all k, then some of the larger ϵ_k values may lead to a group unnecessarily incurring $\frac{1}{\Delta^2}$ dependence (even if using the smallest ϵ would have led to a large gap – again, see Figure 2 in Appendix G). One seemingly natural choice is to scale down each ϵ_k and Δ_k by a constant factor on each iteration, with the smaller of ϵ and Δ using a constant of 2 (and the other constant being chosen to ensure the correct K-th value). However, we leave it for future work as to whether this choice has near-optimality guarantees, or whether better universal choices exist.

4. Lower Bounds

Our final upper bounds on the number of arm pulls, given in Corollary 4 and Corollary 5, have several dependencies of interest, including the number of partitions m, the gaps $\widetilde{\Delta}_{G,i}$, and the multiplicative term $3\epsilon N = O\big(\frac{1}{\epsilon}\log\frac{|\mathcal{G}|}{\delta}\big)$. We discuss some of these dependencies in Section 3.5 and Appendix F.3, and we further discuss the difficulties of obtaining instance-dependent lower bounds in Appendix G.

In addition, we derived the weakened upper bound (22), which can also be viewed as corresponding to (21) in the case that all $\widetilde{\Delta}_{G,i}$ are on the same order as Δ, i.e., a "worst-case" instance. Observe that when $\delta = \frac{\epsilon}{2}$ (recall that we constrain $\delta < \epsilon$), (22) reduces to[7]

$$T(\epsilon, \delta, \Delta) \leqslant \widetilde{O}\left(\frac{|\mathcal{G}|}{\Delta^2 \epsilon^2}\right). \tag{24}$$

Our main result of this section is the following, which states a near-matching $\Omega\big(\frac{1}{\Delta^2\epsilon^2}\big)$ lower bound for worst-case instances.

Theorem 6 (Worst-Case Lower Bound) *For any $\epsilon, \Delta \in \big(0, \frac{1}{4}\big)$, any given number of groups $|\mathcal{G}| \geqslant 2$,[8] any constant value of $\delta \in \big(0, \frac{1}{2}\big)$, and any algorithm for our max-quantile grouped bandit problem with $\alpha = \frac{1}{2}$ guaranteeing (4) with probability at least $1 - \delta$, there exists an instance with Bernoulli rewards such that the time horizon T satisfies*

$$\mathbb{E}[T] \geqslant \Omega\left(\frac{|\mathcal{G}|}{\Delta^2 \epsilon^2}\right). \tag{25}$$

That is, the worst-case upper bound in (24) is tight up to logarithmic factors.

7. The notation $\widetilde{O}(\cdot)$ hides a $\log^2 \frac{|\mathcal{G}|}{\epsilon} + \log \frac{|\mathcal{G}|}{\epsilon} \cdot \log\log \frac{1}{\Delta}$ factor. If we are interested in constant error probability, we could slightly modify our analysis to use two values δ_1 and δ_2 to replace the two occurrences of δ in Lemmas 7 and 8. Then we would only require $\delta_2 < \epsilon$, and the hidden dependence would improve to $\log \frac{|\mathcal{G}|}{\epsilon} \cdot \log\log \frac{1}{\Delta}$.

8. We thank an anonymous reviewer for suggesting that we generalize this result from $|\mathcal{G}| = 2$ to general $|\mathcal{G}| \geqslant 2$.

The proof is given in Appendix E, and is based on a direct analysis of pairwise log-likelihood ratios between suitably-chosen "hard-to-distinguish" instances (having differing optimal groups) with Bernoulli rewards. With $\alpha = \frac{1}{2}$, these instances are roughly defined as follows:

- In all groups except the first, there are only two types of arms, "good" and "bad", with means $\frac{1}{2} + O(\Delta)$ and $\frac{1}{2} - O(\Delta)$ respectively.
- We let group 1 consist entirely of arms with mean $\frac{1}{2}$. The remaining groups are either "good" with a $\frac{1}{2} + O(\epsilon)$ fraction of good arms, or "bad" with a $\frac{1}{2} - O(\epsilon)$ fraction of good arms. We consider one instance where groups $2, \dots, |\mathcal{G}|$ are all bad (and hence group 1 is optimal), as well as $|\mathcal{G}| - 1$ additional instances where only a single group $j \in \{2, \dots, |\mathcal{G}|\}$ is good (and hence group j is optimal).

Then, the rough idea is that for each group indexed by $j \in \{2, \dots, |\mathcal{G}|\}$, we need to consider $\Omega\big(\frac{1}{\epsilon^2}\big)$ arms from the group to determine whether or not it is a good group, but we also need to pull those arms $\Omega\big(\frac{1}{\Delta^2}\big)$ times each to determine whether they are good or bad arms, leading to $\Omega\big(\frac{1}{\Delta^2\epsilon^2}\big)$ arm pulls per group. Our formal proof makes this intuition more precise.

Naturally, it would also be of significant interest to develop instance-dependent lower bounds. In Appendix G, we discuss some difficulties in doing so, including how certain standard change-of-measure techniques (e.g., see (Kaufmann et al., 2016)) appear to be insufficient. Despite this, such techniques do suggest that certain gap quantities may be relevant, and in Appendix G we compare those with the analogous quantities from our upper bound.

5. Conclusion

We have introduced the problem of grouped max-quantile infinite-arm bandits, which comes with a variety of unique challenges compared to its component "ingredients" of infinite-arm bandits and grouped bandits. We introduced a two-step elimination-based algorithm and discussed some evidence for the strength of its instance-dependent upper bound, as well as rigorously proving its optimality up to logarithmic factors in a worst-case sense. For future work, we expect that it should be possible to attain an instance-dependent lower bound that attains the gap terms discussed in Appendix G while also exhibiting the improved ϵ-dependence that was lacking therein.

Acknowledgments

This work was supported by the Singapore National Research Foundation (NRF) under grant number A-0008064-00-00.

References

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53, 2010.

Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Conference on Algorithmic Learning Theory*, pages 3–24. PMLR, 2018.

Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103–2116, 1997.

Thomas Bonald and Alexandre Proutière. Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In *Conference on Neural Information Processing Systems*, 2013.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in X-armed bandits. In *Conference on Neural Information Processing Systems*, 2008.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, 2015.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Quantile-regret minimisation in infinitely many-armed bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2019.

Yahel David and Nahum Shimkin. Infinitely many-armed bandits with unknown value distribution. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014.

Gabillon, Victor, Ghavamzadeh, Mohammad, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Conference on Neural Information Processing Systems*, 2012.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In *Conference on Neural Information Processing Systems*, 2011.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

Jean-Bastien Grill, Michal Valko, Remi Munos, and Remi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Conference on Neural Information Processing Systems*, volume 28, 2015.

Samarth Gupta, Shreyas Chaudhari, Subhojyoti Mukherjee, Gauri Joshi, and Osman Yağan. A unified approach to translate classical bandit algorithms to the structured bandit setting. *IEEE Journal on Selected Areas in Information Theory*, 1(3):840–853, 2020.

Ruitong Huang, Mohammad M Ajallooeian, Csaba Szepesvári, and Martin Müller. Structured best arm identification with fixed confidence. In *Conference on Algorithmic Learning Theory*, 2017.

Kevin Jamieson. CSE599i: Online and Adaptive Machine Learning, Lecture 4. https://courses.cs.washington.edu/courses/cse599i/18wi/resources/lecture4/lecture4.pdf, 2022. Accessed: 2022-09-15.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Conference on Information Sciences and Systems*, pages 1–6, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Kevin Jamieson, Daniel Haas, and Benjamin Recht. The power of adaptivity in identifying statistical alternatives. In *Conference on Neural Information Processing Systems*, 2016.

Anand Kalvit and Assaf Zeevi. Bandits with dynamic arm-acquisition costs. *arXiv preprint arXiv:2110.12118*, 2021.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2012.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing*, 2008.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Haifang Li and Yingce Xia. Infinitely many-armed bandits with budget constraints. In *AAAI Conference on Artificial Intelligence*, 2017.

Subhojyoti Mukherjee, Ardhendu Tripathy, and Robert Nowak. Generalized Chernoff sampling for active learning and structured bandit algorithms. *arXiv preprint arXiv:2012.08073*, 2020.

Ojash Neopane, Aaditya Ramdas, and Aarti Singh. Best arm identification under additive transfer bandits. In *Asilomar Conference on Signals, Systems, and Computers*, 2021.

Wenbo Ren, Jia Liu, and Ness B Shroff. Exploring k out of top ρ fraction of arms in stochastic bandits. In *Conference on Artificial Intelligence and Statistics*, 2019.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Overlapping multi-bandit best arm identification. In *IEEE International Symposium on Information Theory*, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Vincent Y F Tan, Krishna Jagannathan, et al. A survey of risk-aware multi-armed bandits. *arXiv preprint arXiv:2205.05843*, 2022.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Conference on Neural Information Processing Systems*, 2008.

Zhenlin Wang and Jonathan Scarlett. Max-min grouped bandits. In *AAAI Conference on Artificial Intelligence*, 2021.

Mengyan Zhang and Cheng Soon Ong. Quantile bandits for best arms identification. In *International Conference on Machine Learning*, 2021.

Appendix A. Proof of Theorem 2 (Correctness)

In Section 3.2, we introduced Events A and B. We now show them to be sufficient for Algorithm 1 to return a group G satisfying (4), and then show that each of these events happen with probability at least $1-\delta$ (see Lemma 7 and 8). This implies that Algorithm 1 returns a group G satisfying (4) with probability at least $1-2\delta$, as stated in Theorem 2.

Lemma 7 *Event A occurs with probability at least $1-\delta$.*

Proof This follows from union bounds and Chernoff-Hoeffding bounds; see Appendix A.1 for the details. ■

Lemma 8 *There exists a choice of `FiniteArmBQID` (see Algorithm 2 in Appendix C) such that under Assumption 1, Event B occurs with probability at least $1-\delta$.*

Proof This is a corollary of Theorem 13 in Appendix C, whose proof is given in Appendix D. ■

Observe that under the event $A \cap B$, the group $\widehat{G}$ returned in Line 6 of Algorithm 1 satisfies (4), since

$$F_{\widehat{G}}^{-1}(1-\alpha+\epsilon) \geqslant F_{\widehat{H}}^{-1}(1-\alpha) \geqslant \max_{H\in\mathcal{H}} F_H^{-1}(1-\alpha) - \Delta \geqslant \max_{G\in\mathcal{G}} F_G^{-1}(1-\alpha-\epsilon) - \Delta. \quad (26)$$

Combining Lemma 7, Lemma 8, and (26) yields the correctness of Algorithm 1, as stated in Theorem 2.

A.1. Proof of Lemma 7 (Bounding the Probability of Event A)

We show that for a fixed $G \in \mathcal{G}$ and the corresponding finite group H, we have

$$F_G^{-1}(1-\alpha-\epsilon) \leqslant F_H^{-1}(1-\alpha) \leqslant F_G^{-1}(1-\alpha+\epsilon). \tag{27}$$

with probability at least $1-\delta/|\mathcal{G}|$. Then a union bound over the groups $G \in \mathcal{G}$ gives us the desired claim.

The complement of event (27) is equivalent to at least one of these two events happening:

(i) $F_H^{-1}(1-\alpha) > F_G^{-1}(1-\alpha+\epsilon)$, or

(ii) $F_H^{-1}(1-\alpha) < F_G^{-1}(1-\alpha-\epsilon)$.

We will upper-bound the probability of each of Event (i) and (ii) by $\frac{\delta}{2|\mathcal{G}|}$, and then apply a union bound.

Let the number of sampled arms be as in Line 1 of Algorithm 1, i.e., $N = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil$. Note that Event (i) is equivalent to more than $N\alpha$ of the sampled arms being from the top-$(1-\alpha+\epsilon)$ quantile. For $i = 1, \ldots, N$, we let $X_i = 1$ if the i-th sampled arm is from the top-$(1-\alpha+\epsilon)$ quantile, and $X_i = 0$ otherwise. Then $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables with success probability at most $\alpha - \epsilon$, and $X = \sum_{i=1}^N X_i$ is the number of sampled arms in the top-$(1-\alpha+\epsilon)$ quantile, with $\mathbb{E}[X] \leqslant N(\alpha-\epsilon)$. By Hoeffding's inequality, we obtain

$$\mathbb{P}[X \geqslant N\alpha] = \mathbb{P}[X - (N(\alpha-\epsilon)) \geqslant N\epsilon] \leqslant \mathbb{P}[X - \mathbb{E}[X] \geqslant N\epsilon] \leqslant \exp\left(-2N\epsilon^2\right) \leqslant \frac{\delta}{2|\mathcal{G}|}.$$

Similar reasoning shows that the probability of Event (ii) is also bounded above by $\frac{\delta}{2|\mathcal{G}|}$, which completes the proof.

Appendix B. Proof of Corollary 4 (Number of Arm Pulls Without $\mathcal{H}$ Dependence)

B.1. Derivations of Weakened Gaps

Here we provide the derivations of the gap lower bounds stated in (14), (15) and (19). Under Event A, we can lower bound each Δ_H by observing that

$$\begin{aligned}\widetilde{\Delta}_G := \max_{G' \in \mathcal{G}} F_{G'}^{-1}(1-\alpha-\epsilon) - F_G^{-1}(1-\alpha+\epsilon) &\leqslant \max_{H' \in \mathcal{H}} F_{H'}^{-1}(1-\alpha) - F_H^{-1}(1-\alpha) \\ &= F_{H^*}^{-1}(1-\alpha) - F_H^{-1}(1-\alpha) = \Delta_H. \end{aligned} \tag{28}$$

Similarly, we define the following value $\widetilde{\Delta}_0$, which lower bounds Δ_0 under Event A:

$$\widetilde{\Delta}_0 := \max_{G \in \mathcal{G}} F_G^{-1}(1-\alpha-\epsilon) - \max_{G' \in \mathcal{G}, G' \neq G_\epsilon^*} F_{G'}^{-1}(1-\alpha+\epsilon), \tag{29}$$

where $G^*_\epsilon \in \operatorname{argmax}_{G\in\mathcal{G}} F_G^{-1}(1-\alpha+\epsilon)$. Indeed, we have under Event A that

$$\max_{G\in\mathcal{G}} F_G^{-1}(1-\alpha-\epsilon) - \max_{G'\in\mathcal{G}, G'\neq G^*_\epsilon} F_{G'}^{-1}(1-\alpha+\epsilon) \leqslant F_{H^*}^{-1}(1-\alpha) - \max_{H\in\mathcal{H}, H\neq H^*} F_H^{-1}(1-\alpha) = \Delta_0, \tag{30}$$

where the inequality uses the definition of H^* (first term) and the fact that we subtract the second-highest value of a smaller quantity (second term). Note that $\widetilde{\Delta}_0$ can be negative (e.g., when $\Delta_0 = 0$), but this is not a problem since we will later take the maximum with other gaps (see (20)).

It remains to derive (19). For each $0 \leqslant i < \lfloor(1-\alpha)/\epsilon\rfloor - 1$, and for each $j \in S_{H,i}$, we have the following lower bound:

$$\Delta'_{H,j} = \left|\mu_j - F_H^{-1}(1-\alpha)\right| \tag{31}$$
$$= F_H^{-1}(1-\alpha) - \mu_j \tag{32}$$
$$\geqslant F_G^{-1}(1-\alpha-\epsilon) - F_G^{-1}(b_{i+1}), \tag{33}$$

where (33) follows from (7) and the definition of $S_{H,i}$. Likewise, for each $\lfloor(1-\alpha)/\epsilon\rfloor + 1 < i \leqslant m$, and for each $j \in S_{H,i}$, we must have

$$\Delta'_{H,j} = \left|\mu_j - F_H^{-1}(1-\alpha)\right| \tag{34}$$
$$= \mu_j - F_H^{-1}(1-\alpha) \tag{35}$$
$$\geqslant F_G^{-1}(b_i) - F_G^{-1}(1-\alpha+\epsilon). \tag{36}$$

Summarizing the above, we obtain the desired result that $\Delta'_{H,j} \geqslant \widetilde{\Delta}'_{G,i}$ with $\widetilde{\Delta}'_{G,i}$ defined in (20) (the third case therein is trivial).

B.2. Intermediate Results and Proof of Corollary 4 (Number of Arm Pulls)

Recall the lower bound on the gaps introduced in (20), which ensures that every arm j in the same subset $S_{H,i}$ shares a common lower bound $\Delta_{H,j} \geqslant \widetilde{\Delta}_{G,i}$. Summing over the groups $H \in \mathcal{H}$ and the subsets $S_{H,i}$ in (12), we obtain an upper bound of the following form.

Corollary 9 *Under Assumption 1, with probability at least* $1-2\delta$*, the total number of arm pull used by Algorithm 1 satisfies*

$$T(\epsilon,\delta,\Delta) \leqslant \sum_{H\in\mathcal{H}} \sum_{i=1}^{m} \frac{c}{\widetilde{\Delta}^2_{G(H),i}} \log\left(\frac{|\mathcal{H}|N}{\delta} \log \frac{1}{\widetilde{\Delta}^2_{G(H),i}}\right) \cdot |S_{H,i}|, \tag{37}$$

where c *is a universal constant,* m *is as defined in* (16), $\widetilde{\Delta}_{G(H),i}$ *is as defined in* (20), *and* $S_{H,i}$ *is as defined in* (17).

While the bound in (37) does not contain the terms $\Delta_{H,j}$, it is still dependent on the specific realization of H through the terms $|S_{H,i}|$. To remove this remaining dependence on $\mathcal{H}$, we will show that with high probability, $|S_{H,i}| \leqslant 3\epsilon N$ for each H and i. In other words, when drawing arms from G to form H, with high probability, at most $3\epsilon N$ arms are from the interval $[b_i, b_{i+1})$.

Lemma 10 *Let $(\mathcal{G}, \alpha, \epsilon, \Delta, \delta)$ be a valid input of Algorithm 1. Let N be as in Line 1 of Algorithm 1, and let m be as in (16). For each $G \in \mathcal{G}$ and its random arm group $H = H(G)$ generated in Lines 2–3 of Algorithm 1, we partition H into $m+1$ disjoint multisets $\{S_{H,i}\}_{i=0}^{m}$ as defined in (17). Then with probability at least $1-\delta$, we have $|S_{H,i}| \leqslant 3\epsilon N = O\big(\frac{1}{\epsilon}\log\frac{|\mathcal{G}|}{\delta}\big)$ for each $H \in \mathcal{H}$ and for each $i = 0, \ldots, m$.*

Proof This follows easily from Hoeffding's inequality and the union bound; see Appendix B.3 for the details. ■

Combining Theorem 2, Corollary 9 and Lemma 10, we obtain Corollary 4 as desired.

B.3. Proof of Lemma 10 (Partitioning of Arms)

For each $H \in \mathcal{H}$, and $i = 0, 1, \ldots, m$, we let $E_{H,i}$ denote the event that $|S_{H,i}| \leqslant 3\epsilon N$, where we recall the choice

$$N = \left\lceil \frac{1}{2\epsilon^2}\log\frac{2|\mathcal{G}|}{\delta}\right\rceil = \left\lceil \frac{1}{2\epsilon^2}\log\frac{2|\mathcal{H}|}{\delta}\right\rceil. \tag{38}$$

We will show that the complement of each event $E_{H,i}$ occurs with probability at most $\frac{\delta}{(m+1)|\mathcal{H}|}$. Then, a union bound over all $H \in \mathcal{H}$ and $i = 0, \ldots, m$ gives us the desired claim.

Fix an arbitrary $H \in \mathcal{H}$ and an arbitrary $i \in \{0, \ldots, m\}$. By the definition in (17), the complement of event $E_{H,i}$ is equivalent to more than $3\epsilon N$ of the N arms independently sampled uniformly from $[0, 1]$ satisfying $j \in [b_i, b_{i+1})$, where $b_0 := 0$, $b_{m+1} := 1$, and $b_1, \ldots b_m$ are as defined in (18). Each such arm is in $[b_i, b_{i+1})$ with probability at most $b_{i+1} - b_i \leqslant \epsilon$.

For $k = 1, \ldots, N$, we let $X_k = 1$ if the k-th arm sampled from G is from the subset $[b_i, b_{i+1})$, and $X_k = 0$ otherwise. Then $X_1, \ldots, X_N$ are i.i.d. Bernoulli Random Variables with success probability at most ϵ, and $X = \sum_{k=1}^{N} X_k = |S_{H,i}|$, with $\mathbb{E}[X] \leqslant N\epsilon$. By Hoeffding's inequality, we have

$$\mathbb{P}[X > 3N\epsilon] \leqslant \mathbb{P}[X - \mathbb{E}[X] \geqslant 2N\epsilon] \leqslant \exp\left(-8N\epsilon^2\right) \leqslant \frac{\delta^4}{16|\mathcal{H}|^4} < \frac{\delta}{(m+1)|\mathcal{H}|},$$

where the last two steps use the choice of N in (38), along with $\delta^3 < \delta < \epsilon \leqslant \frac{1}{m-1} < \frac{16|\mathcal{H}|^3}{m+1}$.

Appendix C. Details of Subroutine for the Finite-Arm Setting

In this section, we formally define our choice of `FiniteArmBQID` used in Line 5 of Algorithm 1. A complete description of our finite-arm subroutine will be given in Algorithm 2 below. The algorithm maintains a set of active arms and groups (see (44), (45), (46), and Lines 8-10 of Algorithm 2). At each round, the algorithm pulls all arms in the set of active arms, updates their confidence bounds, and eliminates groups that are suboptimal and arms that are "no longer of interest" based on the confidence bounds. When the algorithm identifies that some group satisfies (6) based on the confidence bounds, it terminates and returns that group.

While the algorithm and results in this appendix are used as a stepping stone to the overall guarantees of Algorithm 1, we believe that they are also of interest in their own right.

C.1. Integer-Valued Indexing

Since the number of arms in $\mathcal{H}$ is finite, we *re-index* the set of all arms in $\mathcal{H}$ by $\{1, \dots, n\}$, where

$$n := \left| \bigcup_{H \in \mathcal{H}} H \right| = N|\mathcal{H}| = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{G}|}{\delta} \right\rceil \cdot |\mathcal{G}| \tag{39}$$

recalling the choice of N in Algorithm 1. We will mostly use this integer-valued indexing $j \in \{1, \dots, n\}$ for arms in this section and its associated appendices. If $j \in \{1, \dots, n\}$ is the arm's integer-valued index, $j' \in (0, 1)$ is the arm's index in $[0, 1]$, and H is the arm's group, then we adopt the shorthand notation

$$\Delta_j = \Delta_{H,j'}. \tag{40}$$

We will also slightly abuse notation and write $j \in H$ and $j' \in H$ interchangeably under the two forms of indexing; it will be clear from the context when j is an integer index vs. a continuous index in $[0, 1]$.

C.2. Law of the Iterated Logarithm and Confidence Bounds

As is ubiquitous in MAB problems, our analysis relies on confidence bounds. Despite our distinct objective, our setup still consists of regular arm pulls, and accordingly, we can utilize well-established confidence bounds for stochastic bandits. Many such bounds exist with varying degrees of simplicity vs. tightness, and for concreteness, we focus on the law of the iterated logarithm (as refined by (Kaufmann et al., 2016, Theorem 8); see also (Jamieson et al., 2014, Lemma 3)). This result shows that with high probability, for each arm j and each round index t, the mean reward μ_j is within some confidence interval whose width decreases as $T_j(t)$ increases.

To formally present the result, we introduce and recall some notations and definitions. For each $j \in \{1, \dots, n\}$ and each round index $t \geqslant 1$, let $\widehat{\mu}_{j,T_j(t)}$ denotes the empirical mean of observed rewards of arm j up to round t. We will use the following function $U \colon \mathbb{Z}^+ \times (0, 1) \to \mathbb{R}^+$:

$$U(T, \delta) := \sqrt{\frac{2\log(1/\delta) + 6\log\log(1/\delta) + 3\log\log(eT)}{T}} = \Theta\left(\sqrt{\frac{1}{T}\log\left(\frac{\log T}{\delta}\right)}\right), \tag{41}$$

which describes the width of the confidence interval of a mean reward after T pulls.

Lemma 11 (Anytime confidence bounds (Kaufmann et al., 2016, Theorem 8)) *Given $\delta \in (0, 1)$, under Assumption 1, we have with probability at least $1 - \delta$ that*

$$\mathrm{LCB}_t(j) \leqslant \mu_j \leqslant \mathrm{UCB}_t(j) \quad \textit{for all arms } j \in \{1, \dots, n\} \textit{ and for all rounds } t \geqslant 1, \tag{42}$$

where

$$\mathrm{LCB}_t(j) := \widehat{\mu}_{j,T_j(t)} - U\left(T_j(t), \frac{\delta}{n}\right) \quad \textit{and} \quad \mathrm{UCB}_t(j) := \widehat{\mu}_{j,T_j(t)} + U\left(T_j(t), \frac{\delta}{n}\right) \tag{43}$$

are the lower and upper confidence bounds for arm j at round t.

C.3. Successive Elimination

With the confidence bounds (42) formally defined, we now formally describe how Algorithm 2 maintains the active arms and groups at round t based on confidence bounds. In the following, we write $\mathrm{Q}_{1-\alpha}(X)$ as the $(1-\alpha)$-quantile of a finite multiset X, defined in the same way as (5).

We will work in rounds indexed by $t \geqslant 1$. We first define the set $\mathcal{C}_t$ of candidate *potentially optimal groups* at the beginning of round t, initialized as $\mathcal{C}_1 := \mathcal{H}$ and subsequently updated based on confidence bounds (42) as follows:

$$\mathcal{C}_{t+1} := \left\{ H \in \mathcal{C}_t \;\middle|\; \mathrm{Q}_{1-\alpha}\left(\{\mathrm{UCB}_t(j) : j \in H\}\right) \geqslant \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{LCB}_t(j') : j' \in H'\right\}\right) \; \forall H' \in \mathcal{C}_t \right\}. \tag{44}$$

This definition allows us to eliminate groups that are suboptimal according to the confidence bounds. For each group $H \in \mathcal{C}_t$, we define the set $m_t^{(H)}$ of *potential* $(1-\alpha)$*-quantile arms* of H at the beginning of round t, initialized as $m_1^{(H)} := H$, and subsequently updated based on the confidence bounds as follows:

$$\begin{aligned} m_{t+1}^{(H)} := \Big\{ j \in H \;\big|\; \mathrm{LCB}_t(j) \leqslant \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{UCB}_t(j') : j' \in H\right\}\right) \quad \text{and} \\ \mathrm{UCB}_t(j) \geqslant \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{LCB}_t(j') : j' \in H\right\}\right) \Big\}. \end{aligned} \tag{45}$$

This definition allows us to eliminate arms that are no longer potentially $(1-\alpha)$-quantile arms in H according to confidence bounds (42). Based on (44) and (45), we define the set $\mathcal{B}_t$ of active arms, i.e., the arms that will be pulled in round t, initialized as $\mathcal{B}_1 := \bigcup_{H \in \mathcal{H}} H = \{1, \dots, n\}$, and subsequently updated as follows:

$$\mathcal{B}_{t+1} := \left\{ j \;\middle|\; j \in m_{t+1}^{(H)} \text{ for some } H \in \mathcal{C}_{t+1} \right\}. \tag{46}$$

Recall that we only need to find (with high probability) a group H whose quantile is within Δ of the highest, i.e., satisfies (6). To take advantage of this relaxation, we define the quantity $\Delta^{(t)}$ that tracks the difference between the most optimistic and pessimistic possibilities of $\max_{H \in \mathcal{C}_t} F_H^{-1}(1-\alpha)$ according to confidence bounds at the beginning of round t, initialized as $\Delta^{(1)} := \infty$, and subsequently updated as follows:

$$\Delta^{(t+1)} := \max_{H \in \mathcal{C}_{t+1}} \left\{ \mathrm{Q}_{1-\alpha}\left(\{\mathrm{UCB}_t(j) : j \in H\}\right) \right\} - \max_{H' \in \mathcal{C}_{t+1}} \left\{ \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{LCB}_t(j) : j \in H'\right\}\right) \right\}. \tag{47}$$

Once $\Delta^{(t)} \leqslant \Delta$, each group $H \in \operatorname*{argmax}_{H \in \mathcal{C}_t} \{\mathrm{Q}_{1-\alpha}(\{\mathrm{LCB}_{t-1}(j) : j \in H\})\}$ satisfies (6) provided that the confidence bounds are valid, and so we can simply return any one of them.

With these definitions in place, we now present the pseudo-code for Algorithm 2.

Algorithm 2 Max-Quantile Grouped Finite-Arm Bandit Algorithm (Used as `FiniteArmBQID` in Algorithm 1)

Require: A finite set of finite arm groups $\mathcal{H}$, parameters $\alpha, \Delta, \delta \in (0, 1)$

1: Initialize $t = 1$ and $T_j(0) = 0$ for all $j \in \bigcup_{H \in \mathcal{H}} H = \{1, \ldots, n\}$
2: Set $m_1^{(H)} = H$ for all $H \in \mathcal{H}$; set $\mathcal{C}_1 = \mathcal{H}$, $\Delta^{(1)} = \infty$, and $\mathcal{B}_1 = \bigcup_{H \in \mathcal{H}} H$
3: **while** $|\mathcal{C}_t| > 1$ and $\Delta^{(t)} > \Delta$ **do**
4:     Pull every arm $j \in \mathcal{B}_t$ once
5:     Update $T_j(t) = T_j(t-1) + 1$ if $j \in \mathcal{B}_t$ and $T_j(t) = T_j(t-1)$ otherwise
6:     Compute $U\big(T_j(t), \delta_{j, T_j(t)}\big)$ for every arm $j = \{1, \ldots, n\}$ according to (41)
7:     Compute $\mathrm{LCB}_t(j)$ and $\mathrm{UCB}_t(j)$ for every arm $j = \{1, \ldots, n\}$ according to (43)
8:     Compute $\mathcal{C}_{t+1}$ according to (44)
9:     Compute $m_{t+1}^{(H)}$ for each $H \in \mathcal{C}_{t+1}$ according to (45)
10:     Compute $\mathcal{B}_{t+1}$ and $\Delta^{(t+1)}$ according to (46) and (47)
11:     Increment the round index t by 1
12: Set $\widehat{H} \in \operatorname*{argmax}_{H \in \mathcal{C}_t} \{\mathrm{Q}_{1-\alpha}(\{\mathrm{LCB}_{t-1}(j) : j \in H\})\}$ with uniformly random tie-breaking
13: **return** $\widehat{H}$

Remark 12 *While computational complexity is not our focus, we note that Lines 5–10 can be computed more efficiently than a direct implementation by observing the following:*

- *Every non-eliminated arm has been pulled the same number of times (i.e., $T_j(t) = t$ if $j \in \mathcal{B}_t$), and hence,*
$$U\left(T_j(t), \frac{\delta}{n}\right) = U\left(t, \frac{\delta}{n}\right) \quad \forall j \in \mathcal{B}_t, \tag{48}$$
from which $\mathrm{LCB}_t(j)$ and $\mathrm{UCB}_t(j)$ can readily be computed. Moreover, the fact that every non-eliminated arm has the same confidence width implies that
$$\Delta^{(t+1)} = 2U\left(t, \frac{\delta}{n}\right) \quad \forall t \geqslant 1, \tag{49}$$
with both maxima in (47) being attained by the same group.
- *After eliminating any arm, its UCB and LCB values no longer need to be computed.*

We now state our main result regarding Algorithm 2.

Theorem 13 (Performance Guarantee for Algorithm 2) *Let $(\mathcal{H}, \alpha, \Delta, \delta)$ be a valid input of Algorithm 2. Under Assumption 1, with probability at least $1 - \delta$, Algorithm 2 identifies a group $\widehat{H}$ satisfying (6)* *and uses a number of arm pulls satisfying*

$$T(\delta) \leqslant \sum_{j=1}^{n} \frac{c}{\Delta_j^2} \log\left(\frac{n}{\delta} \log \frac{1}{\Delta_j^2}\right) \tag{50}$$

where c is a universal constant, n is as defined in (39), and Δ_j is as defined in (8).

Proof The proof follows typical steps used in elimination-based algorithms for finite-arm settings (with (Wang and Scarlett, 2021) being perhaps the most similar), but requires care in handling the variety of cases that may occur. See Appendix D for the details. ■

Observe that Theorem 13 is equivalent to the first statement in Theorem 3, with the latter using continuous indexing instead of integer indexing.

Appendix D. Proof of Theorem 13 (Performance Guarantee for Algorithm 2)

Throughout this appendix, we use the integer-valued indexing convention introduced in Appendix C.1, in particular using Δ_j as per (40).

Recalling Lemma 11, it is sufficient to show that when the confidence bounds (43) are valid (with parameter δ), Algorithm 2 identifies a group $\widehat{H}$ satisfying (6) and uses a number of arm pulls satisfying (50). We separate this conditional performance guarantee into two parts: correctness in Lemma 14, and bounding the number of arm pulls in Corollary 16.

Lemma 14 (Conditional Correctness of Algorithm 2) *If the confidence bounds (43) are valid, then Algorithm 2 returns a group $\widehat{H}$ satisfying (6).*

Proof See Appendix D.1. ■

For the conditional bound on arm pulls, we show that if the confidence bounds (43) are valid, then the condition $U\big(T_j(t), \frac{\delta}{n}\big) < \frac{\Delta_j}{4}$ is sufficient to conclude that arm j will not be pulled at rounds $\tau \geqslant t+1$. This condition allows us to find an upper bound on the number of pulls of arm j. Summing over all arms $j \in \{1, \ldots, n\}$ yields the bound.

Lemma 15 *If the confidence bounds (43) are valid, then for each arm $j \in \{1, \ldots, n\}$ and any time t, we have*

$$U\left(T_j(t), \frac{\delta}{n}\right) < \frac{\Delta_j}{4} \implies \left(j \notin \mathcal{B}_{t+1} \text{ or } |\mathcal{C}_{t+1}| = 1 \text{ or } \Delta^{(t+1)} \leqslant \Delta\right)$$

That is, after the round index t satisfies $U\big(T_j(t), \frac{\delta}{n}\big) < \frac{\Delta_j}{4}$, arm j will no longer be pulled in Algorithm 2.

Proof See Appendix D.2. ■

To make the condition on t more explicit, we write

$$\min\left\{t : U\left(T_j(t), \frac{\delta}{n}\right) < \frac{\Delta_j}{4}\right\} = \min\left\{t : U\left(t, \frac{\delta}{n}\right) < \frac{\Delta_j}{4}\right\} \leqslant \frac{c}{(\Delta_j)^2} \log\left(\frac{n}{\delta} \log \frac{1}{(\Delta_j)^2}\right), \tag{51}$$

where the last step holds for some universal $c > 0$ by a standard inversion, e.g., (Jamieson, 2022, p.5). Hence, using Lemma 15 and summing over the arms, we obtain the following.

Corollary 16 (Conditional Bounds on Arm Pulls) *If the confidence bounds (43) are valid, then Algorithm 2 uses a number of arm pulls satisfying (50).*

D.1. Proof of Lemma 14 (Correctness of Algorithm 2)

For brevity, we say that an arm is a $(1-\alpha)$-quantile arm in group H if it has a mean reward of $F_H^{-1}(1-\alpha)$, and we denote an arbitrary such arm by $j_\alpha(H)$.[9] We let $H^* \in \operatorname*{argmax}_{H \in \mathcal{H}} F_H^{-1}(1-\alpha)$ be a single optimal group, breaking ties arbitrarily in the case of non-uniqueness.

We first show that $H^* \in \mathcal{C}_t$ for each round $t \geqslant 1$, i.e., H^* always remains a potentially optimal group. For each fixed $t \geqslant 1$, we let $\mathcal{E}_t$ denote the event that $H^* \in \mathcal{C}_t$. We show by induction that $\mathcal{E}_t$ holds for all $t \geqslant 1$. For $t = 1$, we have $H^* \in \mathcal{H} = \mathcal{C}_1$. We now show the inductive step: When $\mathcal{E}_t$ holds, so does $\mathcal{E}_{t+1}$. For all $H' \in \mathcal{C}_t$, we have

$$\mathrm{Q}_{1-\alpha}\left(\{\mathrm{UCB}_t(j) : j \in H^*\}\right) \geqslant \mathrm{Q}_{1-\alpha}\left(\{\mu_j : j \in H^*\}\right) \tag{52}$$

$$= \mu_{j_\alpha(H^*)} \tag{53}$$

$$\geqslant \mu_{j_\alpha(H')} \tag{54}$$

$$= \mathrm{Q}_{1-\alpha}\left(\left\{\mu_{j'} : j' \in H'\right\}\right) \tag{55}$$

$$\geqslant \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{LCB}_t(j') : j' \in H'\right\}\right), \tag{56}$$

where (52) and (56) follow from the confidence bounds (42)–(43), and (54) uses the definition of H^*. By (56) and the definition of $\mathcal{C}_{t+1}$ (see (44)), we conclude that $H^* \in \mathcal{C}_{t+1}$ as desired.

We now argue that the while-loop of Algorithm 2 will terminate, and the returned group $\widehat{H}$ satisfies (6). The halting criteria of while-loop will eventually be satisfied because the width of confidence intervals satisfies $U(T, \delta) \to 0$ as $T \to \infty$ for any $\delta > 0$ (see (41)). If the while-loop of Algorithm 2 terminates because $|\mathcal{C}_t| = 1$, then $\mathcal{C}_t = \{H^*\}$. It trivially follows that the returned group $\widehat{H} = H^*$ satisfies (6). On the other hand, if the while-loop terminates because $\Delta^{(t)} \leqslant \Delta$ for some $t \geqslant 2$, then for an

9. The use of infimum in (5) ensures that such an arm always exists.

arbitrary $\widehat{H} \in \underset{H \in \mathcal{C}_t}{\operatorname{argmax}} \{\mathrm{Q}_{1-\alpha}(\{\mathrm{LCB}_{t-1}(j) : j \in H\})\}$, we have

$$\begin{aligned}
F_{\widehat{H}}^{-1}(1-\alpha) &= \mathrm{Q}_{1-\alpha}\big(\{\mu_j : j \in \widehat{H}\}\big) && (57)\\
&\geqslant \mathrm{Q}_{1-\alpha}\big(\{\mathrm{LCB}_{t-1}(j) : j \in \widehat{H}\}\big) && (58)\\
&= \max_{H \in \mathcal{C}_t} \{\mathrm{Q}_{1-\alpha}(\{\mathrm{LCB}_{t-1}(j) : j \in H\})\} && (59)\\
&= \max_{H \in \mathcal{C}_t} \{\mathrm{Q}_{1-\alpha}(\{\mathrm{UCB}_{t-1}(j) : j \in H\})\} - \Delta^{(t)} && (60)\\
&\geqslant \max_{H \in \mathcal{C}_t} \{\mathrm{Q}_{1-\alpha}(\{\mu_j : j \in H\})\} - \Delta && (61)\\
&= \mathrm{Q}_{1-\alpha}(\{\mu_j : j \in H^*\}) - \Delta && (62)\\
&= \max_{H \in \mathcal{H}} F_H^{-1}(1-\alpha) - \Delta, && (63)
\end{aligned}$$

where (58) and (61) follow from confidence bounds and $\Delta^{(t)} \leqslant \Delta$, (60) follows from (47), and (62) follows from $H^* \in \mathcal{C}_t$. Hence, $\widehat{H}$ satisfies (6) in both cases.

D.2. Proof of Lemma 15 (Sufficient Conditions for No Longer Being Pulled)

We first present a useful auxiliary lemma. Observe that if $U\big(T_j(t), \frac{\delta}{n}\big) < \frac{\Delta_j}{4}$ and the confidence bounds are valid, we have

$$\max\Big\{\mu_j - \mathrm{LCB}_t(j), \mathrm{UCB}_t(j) - \mu_j\Big\} \leqslant \mathrm{UCB}_t(j) - \mathrm{LCB}_t(j) = 2\,U\left(T_j(t), \frac{\delta}{n}\right) < \frac{\Delta_j}{2}. \quad (64)$$

We use this observation to prove the following.

Lemma 17 *Let arm $j \in \{1, \cdots, n\}$ be arbitrary and let H be the group containing j. Moreover, suppose that the confidence bounds are valid. If the round index $t \geqslant 1$ satisfies $U\big(T_j(t), \frac{\delta}{n}\big) < \frac{\Delta_j}{4}$, then we have*

$$F_H^{-1}(1-\alpha) + \frac{\Delta_j}{2} > \mathrm{Q}_{1-\alpha}(\{\mathrm{UCB}_t(k) : k \in H\}), \quad (65)$$

and

$$F_H^{-1}(1-\alpha) - \frac{\Delta_j}{2} < \mathrm{Q}_{1-\alpha}(\{\mathrm{LCB}_t(k) : k \in H\}). \quad (66)$$

Proof *To show (65), it is sufficient to show that for any arm $k \in H$, the following implication is true:*

$$\mu_k \leqslant F_H^{-1}(1-\alpha) \implies \mathrm{UCB}_t(k) < F_H^{-1}(1-\alpha) + \frac{\Delta_j}{2}. \quad (67)$$

This is because there are at least a $(1-\alpha)$ fraction of arms in H satisfying $\mu_k \leqslant F_H^{-1}(1-\alpha)$, and each such arm k further satisfies $\mathrm{UCB}_t(k) < F_H^{-1}(1-\alpha) + \frac{\Delta_j}{2}$ under (67), which immediately gives (65). Likewise, for (66), it is sufficient to show that each arm in the set $\big\{k \in H : \mu_k \geqslant F_H^{-1}(1-\alpha)\big\}$ satisfies $F_H^{-1}(1-\alpha) - \frac{\Delta_j}{2} < \mathrm{LCB}_t(k)$. We give a proof for (67) which yields (65), and omit the similar details that yield (66).

For each arm $k \in H$ satisfying $\mu_k \leqslant F_H^{-1}(1-\alpha)$, we have by the definitions (8) and (11) that

$$\Delta_k = \max\left\{\Delta, \Delta_H, \Delta_0, F_H^{-1}(1-\alpha) - \mu_k\right\}. \tag{68}$$

(Recall the integer indexing convention in (40).) Moreover, we have by the same definitions that

$$\Delta_j = \max\left\{\Delta, \Delta_H, \Delta_0, |\mu_j - F_H^{-1}(1-\alpha)|\right\}. \tag{69}$$

We consider two cases: (i) $\Delta_j \geqslant \Delta_k$; and (ii) $\Delta_k > \Delta_j$. For the first case, we have

$$F_H^{-1}(1-\alpha) + \frac{\Delta_j}{2} \geqslant \mu_k + \frac{\Delta_k}{2} > \mathrm{UCB}_t(k),$$

where the first inequality follows from the assumptions on μ_k and Δ_k, and the second inequality follows from (64). For the second case, we must have $\Delta_k = F_H^{-1}(1-\alpha) - \mu_k$ by (68) and (69), and so

$$F_H^{-1}(1-\alpha) + \frac{\Delta_j}{2} = \mu_k + \Delta_k + \frac{\Delta_j}{2} \geqslant \mu_k + \Delta_k > \mathrm{UCB}_t(k),$$

where the last inequality follows from (64). Combining the two cases gives us (67) as desired. ■

Proof [Proof of Lemma 15] Let arm $j \in \{1, \cdots, n\}$ be arbitrary, and let H be the group containing j. If $j \notin \mathcal{B}_t$, then we also have $j \notin \mathcal{B}_{t+1}$, and we are done. Therefore, we may assume without loss of generality that $j \in \mathcal{B}_t$. Likewise, we assume that the while-loop of Algorithm 2 has not terminated yet. We consider four cases for Δ_j:

(i) $\Delta_j = \Delta'_j$

(ii) $\Delta_j = \Delta$

(iii) $\Delta_j = \Delta_H$

(iv) $\Delta_j = \Delta_0 > \Delta_H$,

and show that in each case, at least one of the following happens:

(a) $j \notin \mathcal{B}_{t+1}$, that is, j is eliminated at the end of round t;

(b) $|\mathcal{C}_{t+1}| = 1$, that is, the first condition of the while-loop termination is satisfied;

(c) $\Delta^{t+1} \leqslant \Delta$, that is, the second condition of the while-loop termination is satisfied.

Case (i): $\Delta_j = \Delta'_j$. By the definition of Δ'_j (see (11)), arm j is not a $(1-\alpha)$-quantile arm in group H, which suggests that it should not be included in $m_{t+1}(H)$. We will show that, indeed, arm j is no longer a potential $(1-\alpha)$-quantile arm in H in round $t+1$, i.e., $j \notin m_{t+1}(H)$, and so $j \notin \mathcal{B}_{t+1}$.

We consider two sub-cases: $\mu_j > \mu_{j_\alpha(H)}$ and $\mu_j < \mu_{j_\alpha(H)}$. If $\mu_j > \mu_{j_\alpha(H)}$, we have

$$\begin{aligned}
\mathrm{LCB}_t(j) &> \mu_j - \frac{\Delta_j}{2} &(70)\\
&= \mu_j - \frac{\Delta'_j}{2} &(71)\\
&= \left(\mu_{j_\alpha(H)} + \mu_j - \mu_{j_\alpha(H)}\right) - \frac{\mu_j - \mu_{j_\alpha(H)}}{2} &(72)\\
&= \mu_{j_\alpha(H)} + \frac{\mu_j - \mu_{j_\alpha(H)}}{2} &(73)\\
&= \mu_{j_\alpha(H)} + \frac{\Delta'_j}{2} &(74)\\
&\geqslant \mathrm{Q}_{1-\alpha}\left(\{\mathrm{UCB}_t(j) : j \in H\}\right), &(75)
\end{aligned}$$

where (70) uses (64), (75) uses (65), and both (72) and (74) use the definition of Δ'_j. From (45) and (75), we have that $j \notin m_{t+1}(H)$, hence $j \notin \mathcal{B}_{t+1}$. The case for $\mu_j < \mu_{j_\alpha(H)}$ is similar, which leads to $\mathrm{UCB}_t(j) < \mathrm{Q}_{1-\alpha}\left(\{\mathrm{LCB}_t(j') : j' \in H\}\right)$, and so $j \notin \mathcal{B}_{t+1}$.

Case (ii): $\Delta_j = \Delta$. The idea of this case is to show that each remaining group $H' \in \mathcal{C}_t$ already satisfies (6), meaning that the while-loop should be terminated. Indeed, we have

$$\Delta^{(t+1)} = 2U\left(t, \frac{\delta}{n}\right) = 2\,U\left(T_j(t), \frac{\delta}{n}\right) < \frac{\Delta_j}{2} = \frac{\Delta}{2} \leqslant \Delta, \tag{76}$$

where the first two equalities follow from Remark 12, and the first inequality follows from (64). Therefore, the while-loop of Algorithm 2 is terminated and arm j is no longer pulled after round t.

Case (iii): $\Delta_j = \Delta_H$. By the definition of Δ_H (see (9)), group H is not an optimal group, which suggests that group H should not be included in $\mathcal{C}_{t+1}$. We will show that, indeed, $H \notin \mathcal{C}_{t+1}$, and so $j \notin \mathcal{B}_{t+1}$.

Let $H^* \in \operatorname*{argmax}_{H \in \mathcal{H}} F_H^{-1}(1-\alpha)$ be an optimal group, and $j_\alpha(H^*)$ be its $(1-\alpha)$-quantile arm. By these definitions, $\Delta_{j_\alpha(H^*)} = \max\{\Delta, \Delta_0\} \leqslant \Delta_j = \Delta_H$. Then, we have

$$\begin{aligned}
\mathrm{Q}_{1-\alpha}\left(\{\mathrm{LCB}_t(k) : k \in H^*\}\right) &> \mu_{j_\alpha(H^*)} - \frac{\max\{\Delta, \Delta_0\}}{2} &(77)\\
&\geqslant \mu_{j_\alpha(H^*)} - \frac{\Delta_H}{2} &(78)\\
&= \mu_{j_\alpha(H)} + \frac{\Delta_H}{2} &(79)\\
&= \mu_{j_\alpha(H)} + \frac{\Delta_j}{2} &(80)\\
&\geqslant \mathrm{Q}_{1-\alpha}\left(\{\mathrm{UCB}_t(j) : j \in H\}\right), &(81)
\end{aligned}$$

where (77) uses (66) with H^* and $j_\alpha(H^*)$; (78) uses the assumption that $\Delta_H = \Delta_j \geqslant \max\{\Delta, \Delta_0\}$; (79) uses the definition of Δ_H; and (81) uses (65) with H and j; Then, (44) and (81) imply that $H \notin \mathcal{C}_{t+1}$ as desired.

Case (iv): $\Delta_j = \Delta_0 > \Delta_H$. By the definition of Δ_0 (see (10)), group H must be the unique optimal group, which suggests that $\mathcal{C}_{t+1} = \{H\}$. We will show that, indeed, all other groups H', which are non-optimal, are not included in $\mathcal{C}_{t+1}$, i.e., $\mathcal{C}_{t+1} = \{H\}$. Then the while-loop of Algorithm 2 is terminated because the criterion $|\mathcal{C}_{t+1}| > 1$ no longer holds.

Let H' be an arbitrary suboptimal group. Then

$$\Delta_{H'} = \mu_{j_\alpha(H)} - \mu_{j_\alpha(H')} \geqslant \Delta_0 > 0. \tag{82}$$

Furthermore, its $(1-\alpha)$-quantile arm $j_\alpha(H')$ has gap $\Delta_{\mu_{j_\alpha(H')}} = \max\{\Delta_{H'}, \Delta, \Delta_0\} = \Delta_{H'}$ (note that $\Delta \leqslant \Delta_0$ in the current case). Then, we have

$$\begin{aligned}
\mathrm{Q}_{1-\alpha}\left(\{\mathrm{LCB}_t(j) : j \in H\}\right) &> \mu_{j_\alpha(H)} - \frac{\Delta_0}{2} && (83)\\
&\geqslant \mu_{j_\alpha(H)} - \frac{\mu_{j_\alpha(H)} - \mu_{j_\alpha(H')}}{2} && (84)\\
&= \mu_{j_\alpha(H')} + \frac{\mu_{j_\alpha(H)} - \mu_{j_\alpha(H')}}{2} && (85)\\
&= \mu_{j_\alpha(H')} + \frac{\Delta_{H'}}{2} && (86)\\
&\geqslant \mathrm{Q}_{1-\alpha}\left(\left\{\mathrm{UCB}_t(j') : j' \in H'\right\}\right), && (87)
\end{aligned}$$

where (83) uses (66); (84) and (86) use (82); and (87) uses (65) with H' and $j_\alpha(H')$. Then, (44) and (87) imply that $H' \notin \mathcal{C}_{t+1}$ for all suboptimal groups H'. ∎

Appendix E. Proof of Theorem 6 (Worst-Case Lower Bound)

We consider $\alpha = \frac{1}{2}$, so that the problem is one of best median identification; other values of $\alpha \in (0,1)$ can be handled with only minor changes. It is useful to first handle the case of two groups before handling the general case.

E.1. Proof for the Two-Group Case

For $|\mathcal{G}| = 2$, we construct two instances, one with group 1 being optimal and one with group 2 being optimal, and with the suboptimal group failing to satisfy (4) in both cases. Then, we will show that the two instances are hard to distinguish given the rewards unless $T \geqslant \Omega\left(\frac{1}{\Delta^2\epsilon^2}\right)$.

Reduction to a binary decision problem. We specialize to a simple setting in which there are two types of arms ("good" and "bad"); this specialization is analogous to how the binary-valued problem in (Jamieson et al., 2016) is a special case of the quantile-based problem with general reservoir distributions in (Aziz et al., 2018).[10] Suppose that the arm distributions are Bernoulli, and that group 2's arms have two possible means, $q = \frac{1+\Delta}{2}$ and $\bar{q} = 1 - q = \frac{1-\Delta}{2}$. We call these *good arms* and *bad arms* respectively. Then, the two instances are defined as follows:

10. The difference in the two settings is highlighted by the $\frac{1}{\Delta^2\alpha}$ scaling in (Jamieson et al., 2016), compared to $\frac{1}{\Delta^2\epsilon^2}$ in our setting.

- (Good instance) In group 2, the reservoir distribution places probability mass $p = \frac{1+\epsilon}{2}$ on the good arms, and mass $\bar{p} = 1 - p = \frac{1-\epsilon}{2}$ on bad arms.
- (Bad instance) In group 2, the reservoir distribution places probability mass p on the bad arms, and $\bar{p}$ on the good arms.

In both instances, in group 1, the reservoir distribution places probability mass 1 on arms with mean $\frac{1}{2}$. Hence, group 2 has the higher median in the good instance, and the lower median in the bad instance. Observe that in these instances, up to constant rescaling of ϵ and Δ (e.g., changing them by a factor of 4), we find that the suboptimal group fails to satisfy (4). Since the theorem ignores constant factors, we can ignore this scaling and proceed with ϵ and Δ as above.

Now, if the bandit algorithm knows that the arms and rewards are produced by one of the two instances, the problem is simply reduced to determining whether the instance is good or bad using arms from group 2 alone. (Arms from group 1 provide no information for distinguishing the instances.) Accordingly, all arms mentioned in the subsequent analysis are those from group 2.

In the following, we assume that the bandit algorithm is deterministic. We will derive a lower bound that holds when the good and bad instances each occur with probability $\frac{1}{2}$, and the same lower bound then holds for randomized algorithms by Yao's minimax principle. In addition, it suffices to prove (25) when the target error probability $\delta \in \left(0, \frac{1}{2}\right)$ is an arbitrary fixed positive value (e.g., 0.49). This is because a success probability of 0.51 (say) could be amplified to any value in $\left(\frac{1}{2}, 1\right)$ by independently repeating the algorithm a constant number of times and taking a majority vote, amounting to only a constant factor increase in the number of arm pulls. Finally, we may assume that the time horizon T is fixed, because attaining $\mathbb{E}[T] \leqslant T^*$ implies attaining $T \leqslant CT^*$ with probability at least $1 - \frac{1}{C}$ by Markov's inequality, where C can be arbitrarily large.

Analysis of a single arm. Consider an arbitrary time step of the algorithm and an arbitrary arm that has been pulled some number of times to produce a vector $\mathbf{X}$ of i.i.d. rewards. We first study the likelihood ratio

$$L(\mathbf{X}) = \frac{\mathbb{P}_2[\mathbf{X}]}{\mathbb{P}_1[\mathbf{X}]}, \tag{88}$$

where $\mathbb{P}_2[\cdot]$ (respectively, $\mathbb{P}_1$) denotes probability under the good (respectively, bad) instance. (We will similarly use the notation $\mathbb{E}_1[\cdot]$ and $\mathbb{E}_2[\cdot]$ later.) By our assumption of independent Bernoulli arms, a direct calculation gives that when

$$(\text{\#1s in } \mathbf{X}) - (\text{\#0s in } \mathbf{X}) = d, \tag{89}$$

the likelihood ratio $L(\mathbf{X})$ is given by the following function:

$$f(d) = \frac{pq^d + \bar{p}\bar{q}^d}{\bar{p}q^d + p\bar{q}^d}. \tag{90}$$

For brevity, we refer to d as the *score*.

The following lemmas are central to our analysis, and together they show that the next average of $f(\cdot)$ behaves similarly under the two instances (conditioned on the history so far). Since the proofs are rather technical, they are deferred to Appendices E.3 and E.4.

Lemma 18 *Under the bad instance, for a given arm, let $\mathcal{H}^{(d)}$ be any history of arm pulls giving a score d in (89), and let d' be a random variable indicating the updated score following one additional arm pull after $\mathcal{H}^{(d)}$. Then, we have*

$$\mathbb{E}_1[f(d')|\mathcal{H}^{(d)}] = f(d). \tag{91}$$

Lemma 19 *Under the good instance, for a given arm, let $\mathcal{H}^{(d)}$ be any history of arm pulls giving a score d in (89), and let d' be a random variable indicating the updated score following one additional arm pull after $\mathcal{H}^{(d)}$. Then, we have*

$$\mathbb{E}_2[f(d')|\mathcal{H}^{(d)}] = f(d) + O(\Delta^2\epsilon^2). \tag{92}$$

Analysis of all arms. Suppose that after some number t of arm pulls, a history $\mathcal{H}_t$ has been observed consisting of N arms that have been pulled, with associated scores $d_1, \dots, d_N$. Based on the history, the algorithm selects an arm i^*, leading to (randomly) updated scores $d'_1, \dots, d'_N$ (with $d_j = d'_j$ for all $j \neq i^*$). We claim that

$$\mathbb{E}_2\Big[\prod_{i=1}^{N} f(d'_i)\,\Big|\,\mathcal{H}_t\Big] \leqslant (1 + \mathcal{O}(\epsilon^2\Delta^2)) \prod_{i=1}^{N} f(d_i). \tag{93}$$

To see this, we note that i^* is deterministic given $\mathcal{H}_t$. Hence, we obtain (93) from Lemma 19; the values of d_j for $j \neq i^*$ do not change, so we can factorize their terms out on both sides in (93). Note also that since the arms are independent and the (good) instance is fixed, other arms' rewards do not impact the next reward of arm i^*.

Another point worth mentioning is that a new arm may also be chosen at time t. This is easily captured by the above definitions by increasing N by one and then setting $d_N = 0$ (the score of any not-yet-pulled arm is trivially zero). In fact, without loss of generality, we may assume that T arms are requested at the very start, and accordingly set $N = T$ (though not all of these arms would end up being pulled).

Hence, letting $L_t = \prod_i f(d_i)$ at time t, we have $L_0 = 1$, and if $\mathcal{H}_t$ is the history up to time t, then

$$\mathbb{E}_2[L_{t+1}|\mathcal{H}_t] \leqslant (1 + \mathcal{O}(\epsilon^2\Delta^2))L_t. \tag{94}$$

Applying this recursively, we have after T arm pulls that

$$\mathbb{E}_2[L_T] \leqslant (1 + \mathcal{O}(\epsilon^2\Delta^2))^T \leqslant \exp(\mathcal{O}(T\epsilon^2\Delta^2)). \tag{95}$$

Hence, if $T = O\big(\frac{1}{\epsilon^2\Delta^2}\big)$ with a small enough implied constant, then we have $\mathbb{E}_2[L_T] \leqslant 1.5$. Applying Markov's inequality, it follows that $\mathbb{P}_2[L_T \geqslant 2] \leqslant \frac{3}{4}$. However, when $L_T < 2$, the probability under $\mathbb{P}_1$ of observing the same history is at least half of the corresponding probability under $\mathbb{P}_2$. Given that history, the algorithm can only be correct under at most one of $\mathbb{P}_1$ and $\mathbb{P}_2$. Hence, the failure probability (in selecting between the good and bad instances) is $\Omega(1)$ when $T = O\big(\frac{1}{\epsilon^2\Delta^2}\big)$. In view of our preceding reduction to a binary decision problem, this gives the desired result stated in Theorem 6 for the case that $|\mathcal{G}| = 2$.

E.2. Proof for the General Case

The above analysis of the two-group case can be summarized as showing the following: Let group 1 consist entirely of arms with mean $\frac{1}{2}$, and consider two instances where (i) group 2 contains a fraction $\frac{1-\epsilon}{2}$ of good arms, and (ii) group 2 contains a fraction $\frac{1+\epsilon}{2}$ of good arms. Then any algorithm identifying the optimal group on both of these instances with constant probability (better than random guessing) must have $\mathbb{E}_2[T] \geqslant \Omega\big(\frac{1}{\Delta^2\epsilon^2}\big)$, where the subscript to $\mathbb{E}$ indicates the instance. Moreover, since the problem is symmetric with respect to the two instances (namely, swapping 0 rewards with 1 rewards simply amounts to replacing $\mathrm{Bernoulli}(p)$ arms by $\mathrm{Bernoulli}(1-p)$), such an algorithm must also have $\mathbb{E}_1[T] \geqslant \Omega\big(\frac{1}{\Delta^2\epsilon^2}\big)$.[11]

We now turn to the case of a more general number of groups. We define good and bad arms in the same way as the two-group case (i.e., with means $\frac{1+\Delta}{2}$ and $\frac{1-\Delta}{2}$), but we now generalize the structure of the groups themselves. As before, we let group 1's reservoir distribution place probability mass 1 on arms with mean $\frac{1}{2}$. We then define $|\mathcal{G}|$ different bandit instances according to the proportion of good vs. bad arms in the remaining groups. Specifically, letting p_i denote the proportion of good arms in group $i \in \{2, \dots, |\mathcal{G}|\}$, we define the following:

- In instance 1, we let $p_i = \frac{1-\epsilon}{2}$ for all $i \in \{2, \dots, |\mathcal{G}|\}$.
- In instance j for $j \in \{2, \dots, |\mathcal{G}|\}$, we let $p_j = \frac{1+\epsilon}{2}$, and we let $p_i = \frac{1-\epsilon}{2}$ for all $i \in \{2, \dots, |\mathcal{G}|\} \backslash \{j\}$.

Thus, we clearly have for all $j \in \{1, \dots, |\mathcal{G}|\}$ that group j is the optimal group in instance j. Moreover, up to rescaling of ϵ and Δ, it is the only group that satisfies (4). Thus, under instance j, the algorithm's estimate $\hat{j}$ of the optimal group must equal j with probability at least $1 - \delta$ for some $\delta \in \big(0, \frac{1}{2}\big)$ in accordance with the statement of Theorem 6.

Let $\mathbb{P}_j$ and $\mathbb{E}_j$ denote probability and expectation under instance j. Moreover, for each group i, let T_i be a random variable indicating the total number of pulls of arms from group i. Since $\sum_{i=1}^{|\mathcal{G}|} T_j = T$ almost surely, there must exist a group $j^* \in \{2, \dots, |\mathcal{G}|\}$ such that $\mathbb{E}_1[T_{j*}] \leqslant \frac{\mathbb{E}_1[T]}{|\mathcal{G}|-1}$. We proceed with a reduction to the two-group setting with group 1 and group j^*.

This reduction is based on the observation that all groups except j^* are identical under instance 1 and instance j^* (including group 1), so their arm pulls have no power in distinguishing the two instances. Thus, an algorithm for the $|\mathcal{G}|$-group succeeding on these two instances immediately implies that there exists an algorithm succeeding for the 2-group setting (with only groups 1 and j^*) using an average number of arm pulls given by the quantity $\mathbb{E}_1[T_{j*}]$ mentioned above. Then, the result stated in the first paragraph of this subsection implies that $\mathbb{E}_1[T_{j*}] \geqslant \Omega\big(\frac{1}{\Delta^2\epsilon^2}\big)$, and combining this with $\mathbb{E}_1[T_{j*}] \leqslant \frac{\mathbb{E}_1[T]}{|\mathcal{G}|-1}$ and $|\mathcal{G}| \geqslant 2$ completes the proof of Theorem 6.

11. Viewed differently, we can simply assume without loss of generality that $\mathbb{E}_1[T]$ and $\mathbb{E}_2[T]$ coincide to within a constant factor. To see this, note that if $\mathbb{E}_1[T]$ were much larger than $\mathbb{E}_2[T]$ (say), we could simply run the algorithm up to time $C\mathbb{E}_2[T]$ for some large C, and guess that we are in instance 1 if the algorithm has not terminated. This would increase the error probability under instance 2 by at most $\frac{1}{C}$, which is arbitrarily small.

E.3. Proof of Lemma 18 (Bad Instance)

By a simple application of Bayes' rule, under the bad instance, conditioned on $\mathcal{H}^{(d)}$ with score d, the probability that the arm under consideration is good is

$$\frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d}. \tag{96}$$

As a sanity check, setting $d = 0$ makes this quantity equal to the prior, $\bar{p}$.

From (96) and the fact that good (resp., bad) arms are Bernoulli with mean q (resp., $\bar{q}$), the probability of the next draw returning 1 is given by

$$q\frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d} + \bar{q}\frac{p\bar{q}^d}{\bar{p}q^d + \bar{q}^d}. \tag{97}$$

Therefore, we have

$$\mathbb{E}_1[f(d')|\mathcal{H}^{(d)}] = \left(q\frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d} + \bar{q}\frac{p\bar{q}^d}{\bar{p}q^d + p\bar{q}^d}\right)f(d+1) + \left(\bar{q}\frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d} + q\frac{p\bar{q}^d}{\bar{p}q^d + p\bar{q}^d}\right)f(d-1) \tag{98}$$

$$= \frac{\bar{p}q^{d+1} + p\bar{q}^{d+1}}{\bar{p}q^d + p\bar{q}^d}\frac{pq^{d+1} + \bar{p}\bar{q}^{d+1}}{\bar{p}q^{d+1} + p\bar{q}^{d+1}} + \frac{q\bar{q}(\bar{p}q^{d-1} + p\bar{q}^{d-1})}{\bar{p}q^d + p\bar{q}^d}\frac{pq^{d-1} + \bar{p}\bar{q}^{d-1}}{\bar{p}q^{d-1} + p\bar{q}^{d-1}} \tag{99}$$

$$= \frac{pq^{d+1} + \bar{p}\bar{q}^{d+1} + q\bar{q}(pq^{d-1} + \bar{p}\bar{q}^{d-1})}{\bar{p}q^d + p\bar{q}^d} \tag{100}$$

$$= \frac{(q+\bar{q})(pq^d + \bar{p}\bar{q}^d)}{\bar{p}q^d + p\bar{q}^d} \tag{101}$$

$$= f(d). \tag{102}$$

E.4. Proof of Lemma 19 (Good Instance)

We use similar ideas to the proof of Lemma 18. This time, the conditional probability (given $\mathcal{H}^{(d)}$) of having a good arm is

$$\frac{pq^d}{pq^d + \bar{p}\bar{q}^d}, \tag{103}$$

since in the good instance we swap p and $\bar{p}$. Observe that the difference in the two conditional probabilities ((96) and (103)) can be upper bounded as follows:

$$\frac{pq^d}{pq^d + \bar{p}\bar{q}^d} - \frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d} = \frac{q^d\bar{q}^d(p^2 - \bar{p}^2)}{(pq^d + \bar{p}\bar{q}^d)(\bar{p}q^d + p\bar{q}^d)} \leqslant \frac{q^d\bar{q}^d(p^2 - \bar{p}^2)}{(pq^d)(p\bar{q}^d)} \leqslant 4(p^2 - \bar{p}^2) = 4\epsilon, \tag{104}$$

where the factor of 4 comes from $p \geqslant \frac{1}{2}$, and the last step substitutes $p = \frac{1+\epsilon}{2}$.

Now let A denote the event that the next arm pull returns 1, and let B denote the event that the arm under consideration is good, and observe that

$$\mathbb{P}_2[A|\mathcal{H}^{(d)}] - \mathbb{P}_1[A|\mathcal{H}^{(d)}] \tag{105}$$

$$= q \cdot (\mathbb{P}_2[B|\mathcal{H}^{(d)}] - \mathbb{P}_1[B|\mathcal{H}^{(d)}]) + \bar{q} \cdot (\mathbb{P}_2[B^c|\mathcal{H}^{(d)}] - \mathbb{P}_1[B^c|\mathcal{H}^{(d)}]) \tag{106}$$

$$= (q - \bar{q}) \left(\frac{pq^d}{pq^d + \bar{p}\bar{q}^d} - \frac{\bar{p}q^d}{\bar{p}q^d + p\bar{q}^d} \right) \tag{107}$$

$$\leqslant 4\epsilon\Delta, \tag{108}$$

where we substituted (104) and $q = \frac{1+\Delta}{2}$. It follows that

$$\mathbb{E}_2[f(d')|\mathcal{H}^{(d)}] - \mathbb{E}_1[f(d')|\mathcal{H}^{(d)}]$$

$$= f(d+1)\big(\mathbb{P}_2[A|\mathcal{H}^{(d)}] - \mathbb{P}_1[A|\mathcal{H}^{(d)}]\big) + f(d-1)\big(\mathbb{P}_2[A^c|\mathcal{H}^{(d)}] - \mathbb{P}_1[A^c|\mathcal{H}^{(d)}]\big) \tag{109}$$

$$\leqslant (f(d+1) - f(d-1)) \cdot 4\epsilon\Delta. \tag{110}$$

To bound $f(d+1) - f(d-1)$, consider the function

$$g(x) = \frac{pe^x + \bar{p}}{\bar{p}e^x + p}, \tag{111}$$

and note that

$$f(d) = \frac{pq^d + \bar{p}\bar{q}^d}{\bar{p}q^d + p\bar{q}^d} = g(d \ln(q/\bar{q})). \tag{112}$$

Taking the derivative, we obtain

$$\frac{d}{dx} g'(x) = \frac{e^x(p^2 - \bar{p}^2)}{(\bar{p}e^x + p)^2} \leqslant \frac{e^x(p^2 - \bar{p}^2)}{2p\bar{p}e^x} = \frac{p^2 - \bar{p}^2}{2p\bar{p}} = \mathcal{O}(\epsilon), \tag{113}$$

where the final step uses $p = \frac{1+\epsilon}{2}$ and the assumption $\epsilon \in \left(0, \frac{1}{4}\right)$. Hence, we can bound the difference of interest using a first-order Taylor expansion:

$$f(d+1) - f(d-1) = g((d+1)\ln(q/\bar{q})) - g((d-1)\ln(q/\bar{q})) \tag{114}$$

$$\leqslant 2\ln(q/\bar{q}) \cdot \sup_x g'(x) \tag{115}$$

$$\leqslant 2\ln(q/\bar{q}) \cdot \mathcal{O}(\epsilon) = \mathcal{O}(\epsilon \cdot \Delta), \tag{116}$$

since $q = \frac{1+\Delta}{2}$ and $\Delta \in \left(0, \frac{1}{4}\right)$.

Combining this with (110), and using Lemma 18 for the $\mathbb{E}_1$ term, we obtain (92) as desired.

Appendix F. Comparisons and Connection to Previous Works

F.1. Brief Overview of Two Related Works

Regarding (Wang and Scarlett, 2021) (max-min grouped bandits with finitely-many arms) and (Aziz et al., 2018) (identifying a single $(1-\alpha)$-quantile arm in a non-grouped infinite-arm setting), some similarities and differences to our work are highlighted as follows:

- At a high level, we adopt a two-step approach from (Aziz et al., 2018) of requesting a fixed number of arms (per group) from the reservoir distribution and then running a finite-arm algorithm. Our setting gives rise to fundamental differences (see Appendix F.3), leading to us giving an improved upper bound via a multi-step approach (Section 3.5), as well as requiring distinct techniques for the lower bound (Section 4).
- As we mentioned in Section 3.3, our finite-arm subroutine and its analysis are generally similar to the main algorithm in (Wang and Scarlett, 2021) but with different details (see Appendices C and D). The key distinction is not these details, but rather the challenges of incorporating the finite-arm algorithm into an infinite-arm framework.

F.2. Suboptimality of General Structured Bandit Framework

As mentioned in Section 1.1, various general frameworks for structured bandits have been introduced in the existing literature. Perhaps most notably, our *finite-arm* sub-problem (Appendix C) can be viewed as a special case of best-arm identification in structured bandits, which was studied in (Huang et al., 2017). However, here we discuss how the general-purpose upper bound in (Huang et al., 2017) can be worse than our upper bound in Theorem 13 (also written in continuous-index notation in Theorem 3).

When specialized to our setting, (Huang et al., 2017, Thm. 10) gives gaps of the form $\max\{|\mu_j - c|, \Delta_0/2\}$, with Δ_0 in (10) and c being the mid-point between the best two $(1-\alpha)$-quantiles. We proceed by assuming that $\Delta_0 > 0$, as this follows from a uniqueness assumption made in (Huang et al., 2017).

The weakness is that in suboptimal groups we could have $|\mu_j - c| = 0$ (or ≈ 0) even when Δ'_j in (11) is large, since μ_j could be far above its own group's $(1-\alpha)$-quantile. Moreover, this can happen no matter how small Δ_0 is, if the group under consideration is not among the best two. On the other hand, for arms in the top two groups and arms *below the* $(1-\alpha)$*-quantile* in suboptimal groups, it can be checked that our gaps in (8) match those of (Huang et al., 2017) to within constant factors. Hence, the gaps essentially match for many arms, but not for all arms.

More generally, we are unaware of any follow-up works to (Huang et al., 2017) that could be specialized to obtain our Theorem 13. (We also emphasize that our main results are those for the infinite-arm setting.)

F.3. More Detailed Comparison to Quantile-Based Good Arm Identification

The following discussion concerns our number of arm pulls stated in Corollary 4.

In (Aziz et al., 2018), a related problem was studied in which there is only one infinite-arm group, and the goal is to identify a single arm in the top-α fraction of arms for some $\alpha \in (0,1)$. Their final sample complexity bears some resemblance to ours, e.g., utilizing gaps that depend on quantities $b_1, b_2, \ldots$ amounting to a similar partitioning to that in Section 3.4. In addition, their algorithm uses a similar two-step procedure to our Algorithm 1, first acquiring "sufficiently many" arms and then running a finite-arm algorithm.

Notably, the partitioning in (Aziz et al., 2018) only requires $m = \frac{1}{\alpha}$, whereas we use $m = \frac{1}{\epsilon}$. Similarly, our term $3\epsilon N$ in (21) leads to a further multiplicative $\frac{1}{\epsilon}$ factor, whereas no analogous term is present in (Aziz et al., 2018). The main reason for these differences is that our problem formulation necessitates stricter conditions on the N sampled arms (per group) in order for the finite-arm algorithm to lead to the desired goal. In particular, in (Aziz et al., 2018) it is only required that at least one arm in the top-α fraction is chosen, whereas in our setting it is required that each group's sample quantile is ϵ-close to the true quantile (see Event A in (7)). The former is obtained with N having $\frac{1}{\alpha}$ dependence, whereas the latter requires $\frac{1}{\epsilon^2}$ dependence, making our $3\epsilon N$ term in (21) much more significant.

The preceding discussion also partially addresses the reason why our m value is higher than that of (Aziz et al., 2018), though in Section 3.5, we discuss how this could be at least partially alleviated by moving from a two-step algorithm to a multi-step algorithm. Furthermore, in Section 4, we show that our upper bound is tight up to logarithmic factors in a "worst-case" sense (roughly amounting to all gaps equaling a common value Δ) when $|\mathcal{G}| = 2$ and $\delta = \Theta(1)$, thus justifying our larger choice of N compared to (Aziz et al., 2018).

Appendix G. Discussion on Instance-Dependent Lower Bounds

In this appendix, we provide some discussion on instance-dependent lower bounds; these are absent in Section 4, where we provided a worst-case lower bound.

General-purpose change-of-measure techniques. Change-of-measure techniques have proved to be very useful for obtaining instance-dependent lower bounds in diverse bandit problems (Kaufmann et al., 2016). In our setting, however, it appears to be difficult to use existing general-purpose techniques while maintaining the correct ϵ dependence, as we will discuss further below.

Despite this limitation, it is interesting to observe what dependencies on certain gaps could be obtained via this approach. For simplicity, suppose that there are two groups. Let $\mu_{j_\alpha(G_1)}$ and $\mu_{j_\alpha(G_2)}$ be $(1-\alpha)$-quantiles of the two groups, and suppose without loss of generality that $\mu_{j_\alpha(G_1)} > \mu_{j_\alpha(G_2)}$. Then, let $\epsilon_0 \in [0, \alpha]$ be defined such that a fraction $\alpha - \epsilon_0$ of group 2's arms have mean exceeding $\mu_{j_\alpha(G_1)}$. That is, a fraction ϵ_0 of the arms have means between $\mu_{j_\alpha(G_2)}$ and $\mu_{j_\alpha(G_1)}$. See Figure 2 for an illustration (the quantity ϵ_k therein is discussed later).

In the following, we consider the case $\Delta = 0$ for simplicity. Observe that if any fraction exceeding ϵ_0 of group 2's arms were "shifted" to just above $\mu_{j_\alpha(G_1)}$ (and they were all originally below this value), then group 2 would become the best one. The standard change-of-measure technique in (Kaufmann et al., 2016) then indicates that arms in this subset must be pulled a minimal number of times with dependence $\frac{1}{(\mu_{j_\alpha(G_1)}-\mu_{\min})^2}$, where $\mu_{\min}$ is the smallest mean in the partition. One can then form roughly $\frac{1-\alpha+\epsilon_0}{\epsilon_0}$ disjoint partitions (up to rounding) capturing a fraction (marginally above) ϵ_0 of group 2's reservoir distribution each, and apply the same reasoning.

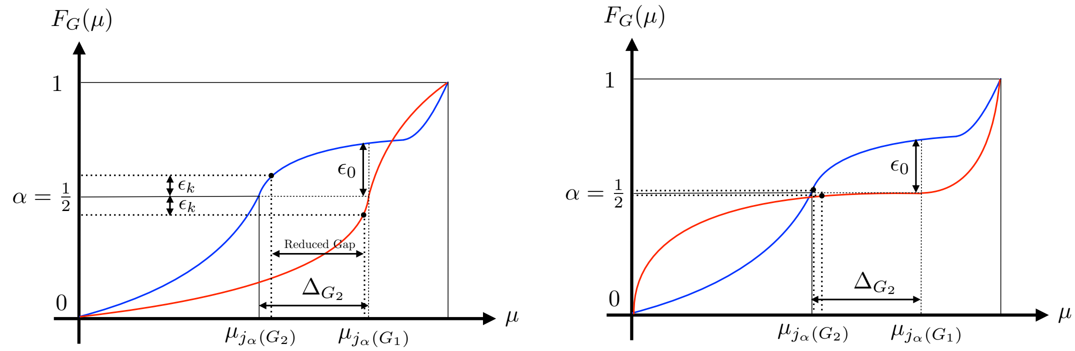

Figure 2: Examples of group reservoir distributions and the associated parameters. In both cases, group 1 is the optimal group, and ϵ_0 takes the same value. In the left figure, the value ϵ_k associated with group 2 is similar to ϵ_0, whereas in the right figure we have $\epsilon_k \ll \epsilon_0$.

The preceding discussion amounts to having gaps of the form $\mu_{j_\alpha(G_1)} - \mu_{\min}$. This is consistent with our use of $\max\{\Delta_H, \Delta'_{H,j}\} = \Theta(\Delta_H + \Delta'_{H,j})$ in our upper bound (though in the upper bound $\mu_{\max}$ would replace $\mu_{\min}$, as we discuss below). Indeed, when j indexes an arm that is in a suboptimal group and not in its top $(1-\alpha)$-quantile, $\Delta_H + \Delta'_{H,j}$ is precisely the difference between $\mu_{H,j}$ and the best group's $(1-\alpha)$-quantile. Similar reasoning can be applied for the number of arm pulls in G^* itself (with a suitably modified choice of ϵ_0).

Comparison of gap terms in upper and lower bounds. In the preceding discussion, the value of ϵ_0 is group-dependent, and it is interesting to compare this value to the quantities $\{\epsilon_k\}_{k=1}^K$ introduced in Section 3.5. As we show in Figure 2, a value of ϵ_k similar in size to ϵ_0 is often large enough to retain a large gap (left example), but in the worst case the required ϵ_k may be significantly smaller than ϵ_0 (right example). The separation between the two is very much dependent on the shapes of the underlying reservoir distributions.

Apart from this key difference, the upper and lower bounds differ for two additional reasons:

- In the upper bound we need to "round" within each partition to lower bound the gap, whereas for the lower bound we need to upper bound the gap. A similar limitation was also present in (Aziz et al., 2018), and as noted therein, the difference becomes less significant as ϵ decreases (or in their setting, as α decreases).
- The lower bound discussed above only characterizes the number of arm pulls on one side of the $(1-\alpha)$-quantile in each group, whereas the upper bound sums over all partitions (i.e., both sides of the quantile). Similar limitations apply to the lower bounds in (Wang and Scarlett, 2021), though in our setting they are alleviated because our α value is fixed in $(0, 1)$ (whereas (Wang and Scarlett, 2021) roughly corresponds to taking $\alpha = 1$).

To summarize the above discussion, the overall gap terms that we introduced share many similar features and properties, but there may still remains significant room for bringing them closer together.

Limitations. Unfortunately, even in cases where near-matching gap terms are attained in the upper and lower bounds, the lower bound approach discussed above would still fail to capture any counterpart to the $3\epsilon N$ factor present in (21) (which has dependence $\frac{1}{\epsilon}$), e.g., only attaining $\frac{1}{\Delta^2\epsilon}$ scaling instead of $\frac{1}{\Delta^2\epsilon^2}$ in (22). We focused on worst-case lower bounds in Theorem 6 because obtaining tight instance-dependent lower bounds appears to be significantly more challenging.