
Exploring k out of Top ρ Fraction of Arms in Stochastic Bandits

Wenbo Ren

Dept. Computer Science & Engineering
The Ohio State University
ren.453@osu.edu

Jia Liu

Dept. Computer Science
Iowa State University
jialiu@iastate.edu

Ness B. Shroff

Dept. Computer Science & Engineering
The Ohio State University
shroff.11@osu.edu

Abstract

This paper studies the problem of identifying any k distinct arms among the top ρ fraction (e.g., top 5%) of arms from a finite or infinite set with a probably approximately correct (PAC) tolerance ϵ . We consider two cases: (i) when the threshold of the top arms' expected rewards is known and (ii) when it is unknown. We prove lower bounds for the four variants (finite or infinite arms, and known or unknown threshold), and propose algorithms for each. Two of these algorithms are shown to be sample complexity optimal (up to constant factors) and the other two are optimal up to a log factor. Results in this paper provide up to $\rho n/k$ reductions compared with the “ k -exploration” algorithms that focus on finding the (PAC) best k arms out of n arms. We also numerically show improvements over the state-of-the-art.

1 INTRODUCTION

Background. Multi-armed bandit (MAB) problems (Berry and Fristedt, 1985) have been studied for decades, and well abstract the problems of decision making with uncertainty. It has been widely applied to many areas, e.g., online advertising (Li et al., 2010), clinical trials (Berry and Eick, 1995), networking (Bubeck and Cesa-Bianchi, 2012; Buccapatnam et al., 2017), and pairwise ranking (Agarwal et al., 2017). In this paper, we focus on stochastic multi-armed bandit. In this setting, each arm of the bandit is assumed to hold a distribution. Whenever the decision maker samples this arm, an independent instance of this distribution is returned. The decision maker

adaptively chooses some arms to sample in order to achieve some specific goals. So far, the majority of works in this area has been focused on minimizing the *regret* (deviation from optimum), (e.g., (Auer et al., 2002; Auer and Ortner, 2010; Garivier and Cappé, 2011; Bubeck and Cesa-Bianchi, 2012; Agrawal and Goyal, 2012; Liu et al., 2018)) that addresses the trade-off between the exploration and exploitation of arms to minimize the regret.

Instead of regret minimization, this paper focuses on pure exploration problems, which aim either (i) to identify one or multiple arms satisfying specific conditions (e.g., with the highest expected rewards) and try to minimize the number of samples taken (e.g., (Mannor and Tsitsiklis, 2004; Kalyan Krishnan and Stone, 2010; Kalyan Krishnan et al., 2012; Cao et al., 2015; Agarwal et al., 2017; Kaufmann and Kalyan Krishnan, 2013; Goschin et al., 2013; Chaudhuri and Kalyan Krishnan, 2017; Aziz et al., 2018)), or (ii) to identify one or multiple best possible arms according to a given criteria within a fixed number of samples (e.g., (Audibert and Bubeck, 2010; Bubeck et al., 2011; Carpentier and Valko, 2015)). In some applications such as product testing (Kohavi et al., 2009; Audibert and Bubeck, 2010; Scott, 2010), before the products are launched, rewards are insignificant, and it is more interesting to explore the best products with the least cost, which suggests the pure exploration setting. This paper focuses on (i) above.

We investigate the problem of identifying any k arms that are in the top ρ fraction of the expected rewards of the arm set. This is in contrast to most works in the pure exploration paradigm that focus on the problem of identifying k best arms of a given arm set. We name the former as the “quantile exploration” (QE) problem, and the latter as the “ k -exploration” (KE) problem, respectively. The motivations of studying the QE problem are as follows: First, in many applications, it is not necessary to identify the best arms, since it is acceptable to find “good enough” arms. For instance, a company wants to hire 100 employees

from more than 10,000 applicants. It may be costly to find the best 100 applicants, and may be acceptable to identify 100 within a certain top percentage (e.g., 5%); Second, theoretical analysis (Kalyanakrishnan et al., 2012; Mannor and Tsitsiklis, 2004) shows that the lower bound on the sample complexity (aka, number of samples taken) of the KE problem depends on n . When the number of arms is extremely large or possibly infinite (e.g., the problem of finding a “good” size from $[10mm, 12mm]$ for a new model of devices), it is infeasible to find the best arms, but may be feasible to find arms within a certain top quantile; Third, by adopting the QE setting, we replace the sample complexity’s dependence on n of the KE problem with k/ρ (Chaudhuri and Kalyanakrishnan, 2017), which can be much smaller, and can greatly reduce the number of samples needed to find “good” arms.

This paper adopts the probably approximately correct (PAC) setting, where an ϵ bounded error is tolerated. This setting can avoid the cases where arms are too close—making the number of samples needed extremely large. The PAC setting has been adopted by numerous previous works (Mannor and Tsitsiklis, 2004; Kalyanakrishnan et al., 2012; Kalyanakrishnan and Stone, 2010; Cao et al., 2015; Goschin et al., 2013; Chaudhuri and Kalyanakrishnan, 2017; Aziz et al., 2018; Kaufmann and Kalyanakrishnan, 2013).

Model and Notations: Let \mathcal{S} be the set of arms. It can be finite or infinite. When \mathcal{S} is finite, let n be its size, and the top ρ fraction arms are simply the top $\lfloor \rho n \rfloor$ arms. If \mathcal{S} is infinite, we assume that the arms’ expected rewards follow some unknown prior identified by an unknown cumulative distribution function (CDF) \mathcal{F} . \mathcal{F} is not necessarily continuous. In this paper, we assume the rewards of the arms are of the same finite support, and normalize them into $[0, 1]$. For an arm a , we use R_a^t to denote the reward of its t -th sample. $(R_a^t, t \in \mathbb{Z}^+)$ are identical and independent. We also assume that the samples are independent across time and arms. For any arm a , let μ_a be its expected reward, i.e., $\mu_a := \mathbb{E}R_a^1$. To formulate the problem, for any $\rho \in (0, 1)$, we define the inverse of \mathcal{F} as

$$\mathcal{F}^{-1}(p) := \sup\{x : \mathcal{F}(x) \leq p\}. \quad (1)$$

The inverse \mathcal{F}^{-1} has the following two properties (2) and (3), where $X \sim \mathcal{F}$ means that X is a random variable following the distribution defined by \mathcal{F} .

$$\mathcal{F}(\mathcal{F}^{-1}(p)) \geq p, \quad (2)$$

$$\mathbb{P}_{X \sim \mathcal{F}}\{X \geq \mathcal{F}^{-1}(p)\} \geq 1 - p. \quad (3)$$

To see (2), by contradiction, suppose $\mathcal{F}(\mathcal{F}^{-1}(p)) < p$. Since $\mathcal{F}(x)$ is right continuous, there exists a number x_1 such that $x_1 > \mathcal{F}^{-1}(p)$ and $\mathcal{F}(x_1) < p$. This implies

that x_1 is in $\{x : \mathcal{F}(x) \leq p\}$, and thus contradicting (1). Define $\mathcal{G}(x) := \mathbb{P}_{X \sim \mathcal{F}}\{X \geq x\}$. Similar to (2), the left continuity of \mathcal{G} implies (3).

In the finite-armed case, an arm a is said to be (ϵ, m) -optimal if $\mu_a + \epsilon \geq \lambda_{[m]}$, where $\lambda_{[m]}$ is defined as the m -th largest expected reward among all arms in \mathcal{S} . In other words, the expected reward of an (ϵ, m) -optimal arm plus ϵ is no less than $\lambda_{[m]}$. The QE problem is to find k distinct (ϵ, m) -optimal arms of \mathcal{S} . We consider both cases where $\lambda_{[m]}$ is known and unknown.

Given a set \mathcal{S} of size n , $k \in \mathbb{Z}^+$ and $\epsilon, \delta \in (0, \frac{1}{2})$, we define the two finite-armed QE problems Q-FK (Quantile, Finite-armed, $\lambda_{[m]}$ Known) and Q-FU (Quantile, Finite-armed, $\lambda_{[m]}$ Unknown) as follows:

Problem 1 (Q-FK). *With known $\lambda_{[m]}$, we want to find k distinct (ϵ, m) -optimal arms with at most δ error probability, and use as few samples as possible.*

Problem 2 (Q-FU). *Without knowing $\lambda_{[m]}$, we want to find k distinct (ϵ, m) -optimal arms with at most δ error probability, and use as few samples as possible.*

In the infinite-armed case, an arm is said to be $[\epsilon, \rho]$ -optimal if its expected reward is no less than $\mathcal{F}^{-1}(1 - \rho) - \epsilon$. Here, we use brackets to avoid ambiguity. To simplify notation, we define $\lambda_\rho := \mathcal{F}^{-1}(1 - \rho)$. An $[\epsilon, \rho]$ -optimal arm is within the top ρ fraction of \mathcal{S} with an at most ϵ error. We consider both cases where λ_ρ is known and unknown. Note that in both cases, we have no knowledge on \mathcal{F} except that λ_ρ is possibly known.

Given a set \mathcal{S} of infinite number of arms, $k \in \mathbb{Z}^+$, and $\rho, \delta, \epsilon \in (0, 1/2)$, we define the two infinite-armed QE problems Q-IK (Quantile, Infinite-armed, λ_ρ Known) and Q-IU (Quantile, Infinite-armed, λ_ρ Unknown).

Problem 3 (Q-IK). *Knowing λ_ρ , we want to find k distinct $[\epsilon, \rho]$ -optimal arms with error probability no more than δ , and use as few samples as possible.*

Problem 4 (Q-IU). *Without knowing λ_ρ , we want to find k distinct $[\epsilon, \rho]$ -optimal arms with error probability no more than δ , and use as few samples as possible.*

2 RELATED WORKS

To our best knowledge, Goschin et al. (2013) was the first one who has focused on the QE problems. They derived the tight lower bound $\Omega(\frac{1}{\epsilon^2}(\frac{1}{\rho} + \log \frac{1}{\delta}))^1$ for the Q-IK problem with $k = 1$. They also provided a Q-IK algorithm for $k = 1$, with sample complexity $O(\frac{1}{\rho \epsilon^2} \log \frac{1}{\delta})$, higher than the lower bound roughly by a $\log \frac{1}{\delta}$ factor. In contrast, our Q-IK algorithm works for all k -values and matches the lower bound.

¹All log, unless explicitly noted, are natural log.

Table 1: Comparison of Previous Works and Ours. All Bounds Are for the Worst Instances.

PROBLEM	WORK	SAMPLE COMPLEXITY
Q-IK	Goschin et al. (2013)	$O\left(\frac{1}{\rho\epsilon^2} \log \frac{1}{\delta}\right)$ for $k = 1$ $\Omega\left(\frac{1}{\epsilon^2} \left(\frac{1}{\rho} + \log \frac{1}{\delta}\right)\right)$ for $k = 1$
	This Paper	$\Theta\left(\frac{k}{\epsilon^2} \left(\frac{1}{\rho} + \log \frac{k}{\delta}\right)\right)$ for $k \in \mathbb{Z}^+$
Q-FK	Goschin et al. (2013)	$O\left(\frac{m}{n\epsilon^2} \log \frac{1}{\delta}\right)$ for $k = 1$
	This Paper	$O\left(\frac{1}{\epsilon^2} \left(n \log \frac{m+1}{m+1-k} + k \log \frac{k}{\delta}\right)\right)$ for $k \leq m \leq n/2$ $\Omega\left(\frac{k}{\epsilon^2} \left(\frac{n}{m} + \log \frac{k}{\delta}\right)\right)$ for $k \leq m \leq n/2$
Q-IU	Chaudhuri et al. (2017)	$O\left(\frac{1}{\rho\epsilon^2} \log^2 \frac{1}{\delta}\right)$ for $k = 1$
	and Aziz et al. (2018)	$\Omega\left(\frac{1}{\rho\epsilon^2} \log \frac{1}{\delta}\right)$ for $k = 1$
	This Paper	$O\left(\frac{1}{\epsilon^2} \left(\frac{1}{\rho} \log^2 \frac{1}{\delta} + k \left(\frac{1}{\rho} + \log \frac{k}{\delta}\right)\right)\right)$ for $k \in \mathbb{Z}^+$ $\Omega\left(\frac{1}{\epsilon^2} \left(\frac{1}{\rho} \log \frac{1}{\delta} + k \left(\frac{1}{\rho} + \log \frac{k}{\delta}\right)\right)\right)$ for $k \in \mathbb{Z}^+$
Q-FU	Chaudhuri et al. (2017)	$O\left(\frac{n}{m\epsilon^2} \log^2 \frac{1}{\delta}\right)$ for $k = 1$ $\Omega\left(\frac{n}{m\epsilon^2} \log \frac{1}{\delta}\right)$ for $k = 1$
	Aziz et al. (2018)	$O\left(\frac{n}{m\epsilon^2} \log^2 \frac{1}{\delta}\right)$ for $k = 1$
	This Paper	$O\left(\frac{1}{\epsilon^2} \left(\frac{n}{m} \log^2 \frac{1}{\delta} + n \log \frac{m+2}{m+2-2k} + k \log \frac{k}{\delta}\right)\right)$ for $2k < m \leq n/2$ $\Omega\left(\frac{1}{\epsilon^2} \left(\frac{n}{m} \log \frac{1}{\delta} + k \left(\frac{n}{m} + \log \frac{k}{\delta}\right)\right)\right)$ for $k \leq m \leq n/2$

Chaudhuri and Kalyanakrishnan (2017) studied the Q-IU and Q-FU problems with $k = 1$. They derived the lower bounds for $k = 1$. In this paper, we generalize their lower bounds to cases where $k > 1$. They also proposed algorithms for these two problems with $k = 1$, and the upper bounds ($O(\frac{1}{\rho\epsilon^2} \log^2 \frac{1}{\delta})$ for Q-IK, $O(\frac{n}{m\epsilon^2} \log^2 \frac{1}{\delta})$ for Q-FK) are the same as ours. For $k > 1$, by simply repeating their algorithms k times and setting error probability $\frac{\delta}{k}$ for each repetition, one can solve the two problems with sample complexity $O(\frac{k}{\rho\epsilon^2} \log^2 \frac{k}{\delta})$ and $O(\frac{n}{mk\epsilon^2} \log^2 \frac{k}{\delta})$, respectively. This paper proposes new algorithms for all k -values with $\log \frac{k}{\delta}$ reductions over the sample complexity.

Aziz et al. (2018) studied the Q-IU problem. They proposed a Q-IK algorithm which is higher than the lower bound proved in this paper by a $\log \frac{1}{\rho\delta}$ factor in the worst case. Under some “good” priors, its theoretical sample complexity can be lower than ours. However, numerical results in this paper show that our algorithm still obtains improvement under “good” priors.

Although the KE problem is not the focus of this paper, we provide a quick overview for comparative perspective. An early attempt on the KE problem was done by Even-Dar et al. (2002), which proposed an algorithm called Median-Elimination that finds an $(\epsilon, 1)$ -optimal arm with probability at least $1 - \delta$ by using at most $O(\frac{n}{\epsilon^2} \log \frac{1}{\delta})$ samples. Mannor and Tsitsiklis (2004); Kalyanakrishnan et al. (2012); Kalyanakrishnan and Stone (2010); Agarwal et al. (2017); Cao et al. (2015); Jamieson et al. (2014); Chen et al.

(2016); Kaufmann and Kalyanakrishnan (2013) studied the KE problem in different settings. Halving algorithm proposed by Kalyanakrishnan and Stone (2010) finds k distinct (ϵ, k) -optimal arms with probability at least $1 - \delta$ by using $O(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$ samples. Kalyanakrishnan and Stone (2010); Kalyanakrishnan et al. (2012); Jamieson et al. (2014); Chaudhuri and Kalyanakrishnan (2017); Aziz et al. (2018); Kaufmann and Kalyanakrishnan (2013) used confidence bounds to establish algorithms that can exploit the large gaps between the arms. In practice, these algorithms are promising, while in theory, their sample complexities may be higher than the lower bound by log factors.

3 LOWER BOUND ANALYSIS

We first establish the Q-FK lower bound.

Theorem 1 (Lower bound for Q-FK). *Given $k \leq m \leq n/2$, $\epsilon \in (0, \frac{1}{4})$, and $\delta \in (0, e^{-8}/40)$, there is a set such that to find k distinct (ϵ, m) -optimal arms of it with error probability at most δ , any algorithm must take $\Omega(\frac{k}{\epsilon^2} (\frac{n}{m} + \log \frac{k}{\delta}))$ samples in expectation.*

Proof Sketch. Mannor and Tsitsiklis (2004, Theorem 13) showed that for the worst instance, finding an $(\epsilon, 1)$ -optimal arm with confidence $1 - \delta$ needs $\Omega(\frac{1}{\epsilon^2} (\frac{n}{m} + \log \frac{1}{\delta}))$ samples in expectation. We show that any Q-FK algorithm for $k = 1$ can be transformed to find $(\epsilon, 1)$ -optimal arms, and derive the desired lower bound for $k = 1$. Then, we construct a series of Q-IK problems with $k = 1$ that match the

lower bound proved above. We show that solving any k of these problems with at most δ total error probability needs at least $\Omega(\frac{k}{\epsilon^2}(\frac{n}{m} + \log \frac{k}{\delta}))$ samples in expectation. Any algorithm that solves the Q-FK problem with parameter k can be transformed to solve the above problems. The desired lower bound follows. \square

By Theorem 1, we establish the Q-IK lower bound.

Theorem 2 (Lower bound for Q-IK). *Given $k, \rho \in (0, \frac{1}{2}]$, $\epsilon \in (0, \frac{1}{4})$, and $\delta \in (0, e^{-8}/40)$, there is an infinite set such that to find k distinct $[\epsilon, \rho]$ -optimal arms of it with error probability at most δ , any algorithm must take $\Omega(\frac{k}{\epsilon^2}(\frac{1}{\rho} + \log \frac{k}{\delta}))$ samples in expectation.*

Proof. By contradiction, suppose there is an algorithm \mathcal{A} that solves all instances of the Q-IK problem by using $o(\frac{k}{\epsilon^2}(\frac{1}{\rho} + \log \frac{k}{\delta}))$ samples in expectation. Choosing $m \geq k(k-1)/\delta$ and $n \geq 2m$, we construct an n -sized set \mathcal{C} that meets the lower bound of the Q-FK problem. By drawing arms from \mathcal{C} with replacement, we can apply \mathcal{A} to it with $\rho = \frac{m}{n}$. Now, we use \mathcal{A} to find k possibly duplicated (ϵ, m) -optimal arms of \mathcal{C} with error probability $\delta/2$. The probability that there is no duplication in these k found arms is at least $\prod_{i=1}^k \frac{m+1-i}{m} \geq 1 - \sum_{i=1}^k \frac{i-1}{m} \geq 1 - \frac{\delta}{2}$. Thus, with probability at least $1 - \delta$, \mathcal{A} finds k distinct (ϵ, m) -optimal arms of \mathcal{C} by $o(\frac{k}{\epsilon^2}(\frac{n}{m} + \log \frac{k}{\delta}))$ samples in expectation, contradicting Theorem 1. The proof is complete. \square

The lower bound for the Q-FU problem directly follows from Theorem 3.3 of (Chaudhuri and Kalyanakrishnan, 2017) and Theorem 1. Theorem 3.3 (Chaudhuri and Kalyanakrishnan, 2017) gives an $\Omega(\frac{n}{m\epsilon^2} \log \frac{1}{\delta})$ lower bounds for $k = 1$. Corollary 3 applies for all k .

Corollary 3 (Lower bound for Q-FU). *Given $k \leq m \leq n/2$, $\epsilon \in (0, 1/\sqrt{32})$, and $\delta \in (0, e^{-8}/40)$, there is a set such that to find k distinct (ϵ, m) -optimal arms with probability at least $1 - \delta$, any algorithm must take $\Omega(\frac{1}{\epsilon^2}(\frac{nk}{m} + k \log \frac{k}{\delta} + \frac{n}{m} \log \frac{1}{\delta}))$ samples in expectation.*

The lower bound for the Q-FU problem directly follows from Corollary 3.4 of (Chaudhuri and Kalyanakrishnan, 2017) and Theorem 2. Corollary 3.4 of Chaudhuri and Kalyanakrishnan (2017) gives an $\Omega(\frac{1}{\rho\epsilon^2} \log \frac{1}{\delta})$ lower bound for $k = 1$. Corollary 4 applies for all k .

Corollary 4 (Lower bound for Q-IU). *Given $k, \rho \in (0, \frac{1}{2}]$, $\epsilon \in (0, 1/\sqrt{32})$, and $\delta \in (0, e^{-8}/40)$, there is an infinite set such that to find k distinct $[\epsilon, \rho]$ -optimal arms with probability at least $1 - \delta$, any algorithm must take $\Omega(\frac{1}{\epsilon^2}(\frac{k}{\rho} + k \log \frac{k}{\delta} + \frac{1}{\rho} \log \frac{1}{\delta}))$ samples in expectation.*

4 ALGORITHMS FOR THE Q-IK PROBLEM

In this section, we present two Q-IK algorithms: AL-Q-IK and CB-AL-Q-IK. ‘‘AL’’ stands for ‘‘algorithm’’

and ‘‘CB’’ stands for ‘‘confidence bounds’’.

A worst case order-optimal algorithm. We first introduce AL-Q-IK. It calls the function ‘‘Median-Elimination’’ (Even-Dar et al., 2002), which finds an $(\epsilon, 1)$ -optimal arm with probability at least $1 - \delta$ by using $O(\frac{|A|}{\epsilon^2} \log \frac{1}{\delta})$ samples. AL-Q-IK is similar to Iterative Uniform Rejection (IUR) (Goschin et al., 2013). At each repetition, IUR draws an arm from \mathcal{S} , performs $\Theta(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ samples on it, and returns it if the empirical mean is large enough. It solves the Q-IK problem with $k = 1$, and its sample complexity is $O(\frac{1}{\epsilon^2\rho} \log \frac{1}{\delta\rho})$. This is higher than the lower bound roughly by a $\frac{1}{\rho} \log \frac{1}{\rho}$ factor (compared with the $\Omega(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ term). The $\frac{1}{\rho} \log \frac{1}{\rho}$ factor is due to the fact that the random arm drawn from \mathcal{S} is $[\epsilon, \rho]$ -optimal with probability ρ (in the worst case). Inspired by their work, we add Lines 2 and 3 to ensure that a_t is $[\epsilon_1, \rho]$ -optimal with probability at least $\frac{1}{2}$. By doing this, we replace the $\frac{1}{\rho} \log \frac{1}{\rho}$ factor by a constant while adding $O(\frac{1}{\rho\epsilon^2})$ samples for each repetition. Repetitions continue until k arms are found, and the number of repetitions is no more than $4k$ in expectation. The choice of n_2 guarantees that for each arm added to Ans , it is $[\epsilon, \rho]$ -optimal with probability at least $1 - \frac{\delta}{k}$. We state its theoretical performance in Theorem 5.

Algorithm 1 AL-Q-IK($\mathcal{S}, k, \rho, \epsilon, \delta, \lambda$)

Input: $\mathcal{S}, k, \rho, \epsilon, \delta$, and $\lambda \leq \mathcal{F}^{-1}(1 - \rho)$;
Initialize: Choose $\epsilon_1, \epsilon_2 > 0$ with $\epsilon_1 + 2\epsilon_2 = \epsilon$;
 $t \leftarrow 0$; $Ans \leftarrow \emptyset$; $n_1 \leftarrow \lceil \frac{1}{\rho} \log 3 \rceil$; $n_2 \leftarrow \lceil \frac{1}{2\epsilon_2^2} \log \frac{k}{\delta} \rceil$;
 $\triangleright \epsilon_1, \epsilon_2 = \Omega(\epsilon)$, Ans stores the chosen arms;

- 1: **repeat** $t \leftarrow t + 1$;
- 2: Draw n_1 arms from \mathcal{S} , and form set A_t ;
- 3: arm $a_t \leftarrow \text{Median-Elimination}(A_t, \epsilon_1, \frac{1}{4})$;
- 4: Sample a_t for n_2 times;
- 5: $\hat{\mu}_t \leftarrow$ the empirical mean;
- 6: **if** $\hat{\mu}_t \geq \lambda_\rho - \epsilon_1 - \epsilon_2$ **then**
- 7: $Ans \leftarrow Ans \cup \{a_t\}$;
- 8: **end if**
- 9: **until** $|Ans| \geq k$
- 10: **return** Ans ;

Theorem 5 (Theoretical performance of AL-Q-IK). *With probability at least $1 - \delta$, AL-Q-IK returns k distinct arms having expected rewards no less than $\lambda - \epsilon$. The expected sample complexity is $O(\frac{k}{\epsilon^2}(\frac{1}{\rho} + \log \frac{k}{\delta}))$.*

Proof Sketch. Correctness: Here, we note that $\lambda_\rho \geq \lambda$. At each repetition, n_1 arms are drawn from \mathcal{S} to guarantee that with probability at least $2/3$, the set A_t contains an arm of the top ρ fraction. Then in Line 3, Median-Elimination($A_t, \epsilon_1, \frac{1}{4}$) is called to get an a_t , which is $[\epsilon_1, \rho]$ -optimal with probability at least $\frac{2}{3}(1 - \frac{1}{4}) = \frac{1}{2}$. At Line 5, by Hoeffding’s Inequality,

we can prove that if a_t is $[\epsilon_1, \rho]$ -optimal, $\hat{\mu}_t$ is greater than $\lambda - \epsilon_1 - \epsilon_2$ with probability at least $1 - \frac{\delta}{k}$, and if $\mu_{a_t} \leq \lambda - \epsilon$, $\hat{\mu}_t$ is less than $\lambda - \epsilon_1 - \epsilon_2$ with probability at least $1 - \frac{\delta}{k}$. By some computation, we show that given a_t is added to Ans , $\mu_{a_t} \geq \lambda - \epsilon$ with probability at least $1 - \frac{\delta}{k}$. Thus, with probability at least $1 - \delta$, all arms in Ans having expected rewards $\geq \lambda - \epsilon$. *Sample Complexity:* For each t , a_t is $[\epsilon_1, \rho]$ -optimal with probability at least $\frac{1}{2}$, and if a_t is $[\epsilon_1, \rho]$ -optimal, then with probability at least $1 - \frac{\delta}{k}$, it will be added to Ans . Thus, in the t -th repetition, with probability at least $(1 - \frac{\delta}{k}) \cdot \frac{1}{2} \geq \frac{1}{4}$, one arm is added to Ans . Thus, the algorithm returns after average $4k$ repetitions. In each repetition, Line 3 takes $O(\frac{n_1}{\epsilon^2} \log 4) = O(\frac{1}{\rho \epsilon^2})$ samples, and Line 4 takes $n_2 = O(\epsilon^{-2} \log(k/\delta))$ samples, proving the sample complexity. \square

Remark: The expected sample complexity of Algorithm 1 matches the lower bound proved in Theorem 2. Even for $k = 1$, this result is better than the previous works $O(\frac{1}{\rho \epsilon^2} \log \frac{1}{\delta})$ (Goschin et al., 2013).

Alternative Version Using Confidence Bounds AL-Q-IK is order-optimal for the worst instances, and provides theoretical insights on the Q-IK problem, but in practice, it does not exploit the large gaps between the arms' expected rewards. In this part, we use confidence bounds to establish an algorithm that is not order-optimal for the worst instance but has better practical performance for most instances. Many previous works (Kalyanakrishnan and Stone, 2010; Kalyanakrishnan et al., 2012; Jamieson et al., 2014; Chaudhuri and Kalyanakrishnan, 2017; Aziz et al., 2018) have shown that this type of confidence-bound-based (CBB) algorithms can dramatically reduce the actual number of samples taken in practice. Given an arbitrary arm a with expected reward μ_a , we let $\hat{X}^N(a)$ be its empirical mean after N samples. A function $u(\cdot)$ ($l(\cdot)$) is said to be an upper (lower) δ -confidence bound if it satisfies

$$\mathbb{P}\{u(\hat{X}^N(a), N, \delta) \geq \mu_a\} \geq 1 - \delta, \quad (4)$$

$$\mathbb{P}\{l(\hat{X}^N(a), N, \delta) \leq \mu_a\} \geq 1 - \delta. \quad (5)$$

There are many choices of confidence bounds, e.g., the confidence bounds using Hoeffding's Inequality can be

$$u(\hat{X}^N(a), N, \delta) = \hat{X}^N(a) + \sqrt{\log \delta^{-1} / (2N)}, \quad (6)$$

$$l(\hat{X}^N(a), N, \delta) = \hat{X}^N(a) - \sqrt{\log \delta^{-1} / (2N)}. \quad (7)$$

In this paper, we propose a general algorithm that works for all confidence bounds satisfying (4) and (5). We first introduce PACMaxing (Algorithm 2), an algorithm to find one (ϵ, m) -optimal arm. The idea follows KL-LUCB (Kaufmann and Kalyanakrishnan, 2013), except that it is designed for all confidence bounds and has a budget to bound the number of samples taken.

Adding *budget* prevents the number of samples from blowing up to infinity, and helps establish Algorithm 3.

In PACMaxing, we let $U^t(a) := u(\hat{\mu}^t(a), N^t(a), \delta^{N^t(a)})$ and $L^t(a) := l(\hat{\mu}^t(a), N^t(a), \delta^{N^t(a)})$. For every arm a , PACMaxing guarantees that during the execution of the algorithm, with probability at least $1 - \frac{\delta}{n}$, its expected reward is always between the lower and upper confidence bounds, and thus, is correct with probability at least $1 - \delta$ (see Lemma 6). Lemma 6's proof is similar to that of KL-LUCB (Kaufmann and Kalyanakrishnan, 2013), and is provided in supplementary materials.

Algorithm 2 PACMaxing($A, \epsilon, \delta, budget$)

Input: A an n -sized set of arms; $\delta, \epsilon \in (0, 1)$;

- 1: $\forall s, \delta^s := \frac{\delta}{k_1 n s^\gamma}$, where $\gamma > 1$ and $k_1 \geq 2(1 + \frac{1}{\gamma-1})$;
- 2: $t \leftarrow 0$ (number of sample taken);
- 3: $B(t) \leftarrow \infty$ (stopping index);
- 4: Sample every arm of A once; $t \leftarrow n$;
- 5: $N^t(a) \leftarrow 1, \forall a \in A$; (number of times a is sampled)
- 6: Let $\hat{\mu}^t(a)$ be the empirical mean of a ;
- 7: $a^t \leftarrow \arg \max_a \hat{\mu}^t(a)$;
- 8: $b^t \leftarrow \arg \max_{a \neq a^t} U^t(a)$;
- 9: **while** $B(t) > \epsilon \wedge t \leq budget$ **do**
- 10: Sample a^t and b^t once; $t \leftarrow t + 2$;
- 11: Update $\hat{\mu}^t(a), \hat{\mu}^t(b), N^t(a), N^t(b)$;
- 12: Update a^t and b^t as Lines 7 and 8;
- 13: $B(t) \leftarrow U^t(b^t) - L^t(a^t)$;
- 14: **end while**
- 15: **if** $B(t) \leq \epsilon$ **then return** a^t
- 16: **else return** a random arm
- 17: **end if**

Lemma 6 (Correctness of PACMaxing). *Given sufficiently large budget, PACMaxing returns an $(\epsilon, 1)$ -optimal arm with probability at least $1 - \delta$.*

Lemma 6 does not provide any insight about PACMaxing's sample complexity because it depends on the confidence bounds we choose. For Hoeffding bounds defined by (6) and (7), we give the sample complexity of PACMaxing in Lemma 7. Its proof is relegated to supplementary materials due to space limitation. Here, we define $\Delta_b := \frac{1}{2} \max\{\epsilon, \max_{a \in A} \mu_a - \mu_b\}$ for all arms b .

Lemma 7 (Sample complexity of PACMaxing). *Using confidence bounds (6) and (7), and for budget no less than $3n + \max\{\frac{8n}{\epsilon^2} \log \frac{k_1 n}{\delta}, \frac{8(1+e^{-1})\gamma n}{\epsilon^2} \log \frac{4(1+e^{-1})\gamma}{\epsilon}\}$, with probability at least $1 - \delta$, PACMaxing returns a correct result after $O(\sum_{a \in A} \frac{1}{\Delta_a} \log \frac{n}{\delta \Delta_a})$ samples.*

Using PACMaxing, we establish the CBB version of AL-Q-IK, presented in Algorithm 3. In the algorithm, we let g_0 and g_1 be the corresponding *budget* lower

bounds as in Lemma 7. CB-AL-Q-IK is almost the same as AL-Q-IK, except that it replaces Median-Elimination and the sampling of a_t by PACMaxing.

Algorithm 3 CB-AL-Q-IK($\mathcal{S}, k, \rho, \epsilon, \delta, \lambda$)

Input: $\mathcal{S}, k, \rho, \epsilon, \delta$, and $\lambda \leq \mathcal{F}^{-1}(1 - \rho)$;
Initialize: Choose $\epsilon_1, \epsilon_2 > 0$ with $\epsilon_1 + 2\epsilon_2 = \epsilon$;
 $t \leftarrow 0$; $Ans \leftarrow \emptyset$; $n_1 \leftarrow \lceil \frac{1}{\rho} \log 3 \rceil$;
 $\triangleright \epsilon_1, \epsilon_2 = \Omega(\epsilon)$, Ans stores the chosen arms;
 1: **repeat** $t \leftarrow t + 1$;
 2: Draw n_1 arms from \mathcal{S} , and form set A_t ;
 3: arm $a_t \leftarrow \text{PACMaxing}(A_t, \epsilon_1, 1/4, g_0)$;
 4: c is an arm with constant rewards $\lambda - \epsilon_1 - \epsilon_2$;
 5: $b_t \leftarrow \text{PACMaxing}(\{a_t, c\}, \epsilon_2, \delta/k, g_1)$;
 6: **if** $b_t = a_t$ **then**
 7: $Ans \leftarrow Ans \cup \{a_t\}$;
 8: **end if**
 9: **until** $|Ans| \geq k$
 10: **return** Ans ;

Theorem 8 states the theoretical performance of CB-AL-Q-IK. Its worst case sample complexity is higher than the lower bound and that of AL-Q-IK roughly by a $\log \frac{1}{\rho\epsilon}$ factor. However, its empirical performance could be better (See Section 7 for numerical evidences).

Theorem 8 (Theoretical performance of CB-AL-Q-IK). *With probability at least $1 - \delta$, CB-AL-Q-IK returns k distinct arms having expected rewards no less than $\lambda - \epsilon$. When using confidence bounds (6) and (7), it terminates after at most $O(\frac{k}{\epsilon^2}(\frac{1}{\rho} \log \frac{1}{\rho\epsilon} + \log \frac{k}{\delta\epsilon}))$ samples in expectation.*

Proof. The correctness follows from directly using the same steps as in the proof of Theorem 5. In each repetition, by Lemma 7, the sample complexity of Line 3 is at most $O(\frac{n_1}{\rho\epsilon^2} \log \frac{n_1}{\rho\epsilon}) = O(\frac{1}{\rho\epsilon^2} \log \frac{1}{\rho\epsilon})$, and that of Line 5 is at most $O(\frac{1}{\epsilon^2} \log \frac{k}{\delta\epsilon})$. The ‘‘at most’’ comes from the choice of *budget* in Lemma 7. The algorithm returns after at most $4k$ repetitions in expectation. The desired sample complexity follows. \square

5 ALGORITHMS FOR THE Q-IU PROBLEM

Chaudhuri and Kalyanakrishnan (2017) proposed an $O(\frac{1}{\rho\epsilon^2} \log^2 \frac{1}{\delta})$ sample complexity algorithm for $k = 1$. Performing it for k times with δ/k error probability for each can solve the problem for all k -values. However, this method will yield unnecessary dependency on $\log^2 k$. If we can first estimate the value of λ_ρ , we can use (CB-)AL-Q-IK to solve this problem and replace the quadratic log dependency by $\log k$. We first use LambdaEstimation (LE) to get a ‘‘good’’ estimation of λ_ρ , and then use AL-Q-IK to solve the Q-IU

problem. We note that this idea may perform poorly for small k -values as evaluating λ_ρ can take more samples than finding several $[\epsilon, \rho]$ -optimal arms.

We first present the algorithm LE for estimating λ_ρ in Algorithm 4. LE calls Halving (Kalyanakrishnan and Stone, 2010), which finds k distinct (ϵ, k) -optimal arms of an n -sized set with probability at least $1 - \delta$ by taking $O(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$ samples. Halving₂ is modified from Halving that finds PAC worst arms (defined by adding a minus to all rewards of the arms).

Algorithm 4 LambdaEstimation($\mathcal{S}, \rho, \epsilon, \delta$) (LE)

Input: \mathcal{S} an infinite set of arms; $\rho, \delta, \epsilon \in (0, 1/2)$;
 1: Choose $\epsilon_1, \epsilon_2, \epsilon_3 = \Omega(\epsilon)$ with $\epsilon_1 + \epsilon_2 + 2\epsilon_3 = \epsilon$;
 2: $n_3 \leftarrow \lceil \frac{32}{\rho} \log \frac{5}{\delta} \rceil$; $n_4 \leftarrow \lceil \frac{1}{2\epsilon_3^2} \log \frac{10}{\delta} \rceil$; $m \leftarrow \lfloor \frac{3}{4} \rho n_3 \rfloor$;
 3: Draw n_3 arms from \mathcal{S} , and form A_1 ;
 4: $A_2 \leftarrow \text{Halving}(A_1, m, \epsilon_1, \frac{\delta}{5})$;
 5: $\hat{a} \leftarrow \text{Halving}_2(A_2, 1, \epsilon_2, \frac{\delta}{5})$;
 6: Sample \hat{a} for n_4 times, $\hat{\mu}_0 \leftarrow$ the empirical mean;
 7: **return** $\hat{\lambda} \leftarrow \hat{\mu}_0 - \epsilon_2 - \epsilon_3$;

In LE, we ensure that with probability at least $1 - \frac{2}{5}\delta$, the m -th most rewarding arm of A_1 is in $M := \{a \in \mathcal{S} : \lambda_\rho \leq \mu_a \leq \lambda_{\rho/2}\}$. After calling Halving and Halving₂, we get \hat{a} , whose expected reward is in $[\lambda_\rho - \epsilon_1, \lambda_{\rho/2} + \epsilon_2]$ with probability at least $1 - \frac{4\delta}{5}$. Finally, \hat{a} is sampled for n_4 times, and its empirical mean is in $[\lambda_\rho - \epsilon_1 - \epsilon_3, \lambda_{\rho/2} + \epsilon_2 + \epsilon_3]$ with probability at least $1 - \delta$. Thus, the returned value $\hat{\lambda}$ is in $[\lambda_\rho - \epsilon, \lambda_{\rho/2}]$ with probability at least $1 - \delta$. Detailed proof of Lemma 9 can be found in supplementary materials.

Lemma 9 (Theoretical performance of LE). *After at most $O(\frac{1}{\rho\epsilon^2} \log^2 \frac{1}{\delta})$ samples, LE returns $\hat{\lambda}$ that is in $[\lambda_\rho - \epsilon, \lambda_{\rho/2}]$ with probability at least $1 - \delta$.*

Now, we use LE to establish the Q-IU algorithm AL-Q-IU. Theorem 10 states its theoretical performance.

Algorithm 5 AL-Q-IU($\mathcal{S}, k, \rho, \epsilon, \delta$)

Input: \mathcal{S} infinite; $k \in \mathbb{Z}^+$; $\rho, \delta, \epsilon \in (0, 1/2)$;
 1: $\hat{\lambda} \leftarrow \text{LE}(\mathcal{S}, \rho, 0.2\epsilon, \frac{\delta}{2})$;
 2: **return** AL-Q-IK($\mathcal{S}, k, \frac{\rho}{2}, 0.8\epsilon, \frac{\delta}{2}, \hat{\lambda}$);

Theorem 10 (Theoretical performance of AL-Q-IU). *With probability at least $1 - \delta$, AL-Q-IU returns k distinct $[\epsilon, \rho]$ -optimal arms. With probability at least $1 - \frac{\delta}{2}$, it terminates after $O(\frac{1}{\epsilon^2}(\frac{1}{\rho} \log^2 \frac{1}{\delta} + k(\frac{1}{\rho} + \log \frac{k}{\delta})))$ samples in expectation.*

Proof. With probability at least $1 - \frac{\delta}{2}$, $\hat{\lambda}$ is in $[\lambda_\rho - \frac{\epsilon}{2}, \lambda_{\rho/2}]$. When $\hat{\lambda}$ is in $[\lambda_\rho - \frac{\epsilon}{2}, \lambda_{\rho/2}]$, by Theorem 5, Line 2 takes $O(\frac{k}{\epsilon^2}(\frac{1}{\rho} + \log \frac{1}{\delta}))$ samples in expectation,

and, with probability at least $1 - \frac{\delta}{2}$, all returned arms are $[\epsilon, \rho]$ -optimal. The correctness of AL-Q-IU follows. The desired sample complexity follows by summing up $O(\frac{k}{\epsilon^2}(\frac{1}{\rho} + \log \frac{1}{\delta}))$ and $O(\frac{1}{\epsilon^2} \log^2 \frac{1}{\delta})$ (Lemma 9). \square

Remark: By Corollary 4, AL-Q-IU is sample complexity optimal up to a $\log \frac{1}{\delta}$ factor. When $\log \frac{1}{\delta} = O(k)$, i.e., $\delta \geq e^{-ck}$ for some constant $c > 0$, AL-Q-IU is sample complexity optimal up to a constant factor.

6 FINITE-ARMED ALGORITHMS

In this section, we let \mathcal{S} be a finite-sized set of arms. By drawing arms from it with replacement, these arms can be viewed as drawn from an infinite-sized set. We use $\mathcal{T}(\mathcal{S})$ to denote the corresponding infinite-sized set, and call it the *infinite extension* of \mathcal{S} .

Q-FK. When $k = 1$, obviously, calling AL-Q-IK($\mathcal{T}(\mathcal{S}), 1, \frac{m}{n}, \epsilon, \delta, \lambda_\rho$) can solve the Q-FK problem. When $k > 1$, we can solve the Q-FK problem by repeatedly calling AL-Q-IK($\mathcal{T}(\mathcal{S}), 1, \rho_t, \epsilon, \delta/k, \lambda_\rho$) and updating \mathcal{S} by deleting the chosen arm, where $\rho_t = \frac{m+1-t}{n+1-t}$. We present the algorithm AL-Q-FK (Algorithm for Q-FK) in Algorithm 6, and state the theoretical performance in Theorem 11. The proof is relegated to supplementary materials.

Algorithm 6 AL-Q-FK($\mathcal{S}, m, k, \epsilon, \delta, \lambda$)

Require: \mathcal{S} n -sized, $k \leq m \leq n/2$, $\lambda \leq \lambda_{[m]}$;
Initialize: $Ans \leftarrow \emptyset$; \triangleright stores the chosen arms;
 1: **repeat**
 2: $\mathcal{S}' \leftarrow \mathcal{T}(\mathcal{S} - Ans)$; $\rho \leftarrow \frac{m-|Ans|}{n-|Ans|}$;
 3: $a_t \leftarrow$ AL-Q-IK($\mathcal{S}', 1, \rho, \epsilon, \delta/k, \lambda$);
 4: $Ans \leftarrow Ans \cup \{a_t\}$;
 5: **until** $|Ans| \geq k$
 6: **return** Ans ;

Theorem 11 (Theoretical performance of AL-Q-FK). *With probability at least $1 - \delta$, AL-Q-FK returns k distinct arms having mean rewards at least $\lambda - \epsilon$. Its takes $O(\frac{1}{\epsilon^2}(n \log \frac{m+1}{m+1-k} + k \log \frac{k}{\delta}))$ samples in expectation.*

Remark: If $k \leq cm$ for some constant $c < 1$, $\log \frac{m+1}{m+1-k} \leq \frac{k}{m+1-k} = O(\frac{k}{m})$, and thus, the expected sample complexity becomes $O(\frac{k}{\epsilon^2}(\frac{k}{m} + \log \frac{k}{\delta}))$, meeting the lower bound (Theorem 1). When k is arbitrarily close to m , the Q-FK problem (almost) reduces to the KE problem. The tightest upper bound for the KE problem (with the knowledge of $\lambda_{[k]}$) is $O(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$ (Kalyanakrishnan et al., 2012) to our best knowledge. When k is arbitrary close to m , as $O(\frac{1}{\epsilon^2}(n \log \frac{m+1}{m+1-k} + k \log \frac{k}{\delta})) = O(\frac{1}{\epsilon^2}(n \log k + k \log \frac{k}{\delta}))$, AL-Q-FK is still better than the literature asymptotically.

Q-FU. AL-Q-FU (Algorithm for solving the Q-FU

problem) is presented in Algorithm 7. Its idea follows AL-Q-IU and AL-Q-FK. We only consider the case $k < \frac{m}{2}$. For $k \geq \frac{m}{2}$, it is better to use KE algorithms instead. Corollary 12 states its theoretical performance, which directly follows from Theorems 10 and 11.

Algorithm 7 AL-Q-FU($\mathcal{S}, m, k, \epsilon, \delta$)

Require: \mathcal{S} n -sized; $2k < m \leq n/2$;
 1: $\hat{\lambda} \leftarrow$ LE($\mathcal{T}(\mathcal{S}), \frac{m}{n}, \frac{\epsilon}{2}, \frac{\delta}{2}$);
 2: **return** AL-Q-FU($\mathcal{S}, \lfloor \frac{m}{2} \rfloor, k, \frac{\epsilon}{2}, \frac{\delta}{2}$);

Corollary 12 (Theoretical Performance of AL-Q-FU). *With probability at least $1 - \delta$, AL-Q-FU returns k distinct (ϵ, m) -optimal arms. With probability at least $1 - \frac{\delta}{2}$, the expected number of samples AL-Q-FU takes is at most $O(\frac{1}{\epsilon^2}(\frac{n}{m} \log^2 \frac{1}{\delta} + n \log \frac{m+2}{m+2-2k} + k \log \frac{k}{\delta}))$.*

Remark: By Corollary 3, when $k \leq cm$ for some constant $c \in (0, \frac{1}{2})$, AL-Q-FU is sample complexity optimal up to a $\log \frac{1}{\delta}$ factor. If $\log \frac{1}{\delta} = O(k)$ also holds, i.e., $\delta \geq e^{-ck}$ for some constant $c > 0$, then AL-Q-FU is sample complexity optimal in order sense.

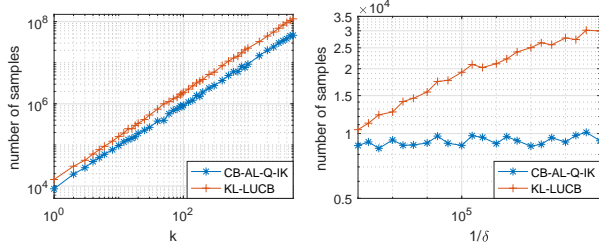
7 NUMERICAL RESULTS

In this section, we illustrate the improvements of our algorithms by running numerical experiments. We present the comparisons of CBB algorithms. The results of the non-CBB algorithms are presented in supplementary materials. Additional numerical results for the finite cases are also presented in supplementary materials. We first compare CB-AL-Q-IK with the literature, and then illustrate the comparison of CB-AL-Q-IU with previous works.

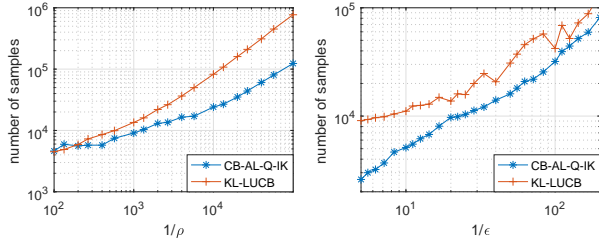
In the simulations, we adopt Bernoulli rewards for all the arms. For fair comparisons, for all CBB-algorithms or versions, we use the KL-Divergence based confidence bounds given by Aziz et al. (2018) with $\gamma = 2$. Every point in every figure is averaged over 100 independent trials. Previous works only considered the case where $k = 1$. In the implementations, for $k > 1$, we repeat them for k times with $\frac{\delta}{k}$ error probability for each repetition.

First, we compare CBB algorithms for the Q-IK problem: CB-AL-Q-IK (choose $\epsilon_1 = 0.75\epsilon$) and (α, ϵ) -KL-LUCB (Aziz et al., 2018) (we name it KL-LUCB in this section). KL-LUCB is almost equivalent to \mathcal{P}_2 (Chaudhuri and Kalyanakrishnan, 2017) with a large enough batch size. The only difference is that they choose different confidence bounds. Here, we note that KL-LUCB does not require the knowledge of λ_ρ . However, we want to show that our algorithm along

with this information can significantly reduce the actual number of samples needed. The priors \mathcal{F} are all Uniform([0,1]). The results are summarized in Figure 1 (a)-(d).



(a) Vary k , $\rho = 0.001, \epsilon = 0.05$, and $\delta = 0.001$. (b) Vary δ , $k = 1$, $\rho = 0.001$, and $\epsilon = 0.05$.



(c) Vary ρ , $k = 1$, $\epsilon = 0.05$, and $\delta = 0.001$. (d) Vary ϵ , $k = 1$, $\rho = 0.001$, and $\delta = 0.001$.

Figure 1: Comparison of CB-AL-Q-IK and KL-LUCB.

It can be seen from Figure 1 that CB-AL-Q-IK performs better than KL-LUCB except two or three points where ρ is large. According to (a), the number of samples CB-AL-Q-IK takes increases slightly slower than KL-LUCB, consistent with the theory that CB-AL-Q-IK depends on $k \log k$ while KL-LUCB depends on $k \log^2 k$. According to (b), we can see that KL-LUCB’s number of samples increases obviously with $\frac{1}{\delta}$, while that of CB-AL-Q-IK is almost independent of δ . The reason is that CB-AL-Q-IK depends on $(\frac{1}{\rho} \log \frac{1}{\rho} + \log \frac{1}{\delta})$ term, and when ρ is small enough, $\log \frac{1}{\delta}$ can be dominated by $\frac{1}{\rho} \log \frac{1}{\rho}$. According to (c), CB-AL-Q-IK takes less samples than KL-LUCB for $\rho < 0.005$, and the gap increases with $\frac{1}{\rho}$. According to (d), CB-AL-Q-IK performs better than KL-LUCB under the given ϵ values.

Second, we compare CB-AL-Q-IU and (α, ϵ) -KL-LUCB. CB-AL-Q-IU is the CBB version of AL-Q-IU by replacing its subroutines by CBB ones. It is designed for large k -values and may not perform well under small k -values, even if it is always in order-sense better or equivalent compared to KL-LUCB. The reason is that its subroutine LE has a large constant factor. However, since the sample complexities of these two algorithms both depend at least linearly on k while that of LE is independent of k ,

when k is large, the influence of (CB-)LE (the CBB version of LE) vanishes, and the improvement of (CB-)AL-Q-IK emerge. The results are summarized in Figure 2. In the implementations, we choose $\epsilon_1 = 0.75\epsilon$ and $\epsilon_2 = 0.1\epsilon$ for algorithm LE, and choose $\epsilon_1 = 0.75\epsilon$ for CB-AL-Q-IK. In Figure 2, the algorithms are tested under a “hard instance” \mathcal{F}_h , where ρ fraction of the arms has expected reward $\frac{1}{2} + 0.55\epsilon$ and the others have $\frac{1}{2} - 0.55\epsilon$. The results are consistent with the theory, and suggest that CB-AL-Q-IK can use much less samples than KL-LUCB does when k is sufficiently large.

We admit that AL-Q-IU may not be practical as it takes 10^8 samples even for $k = 1$, but it also has several contributions: (i) It gives a hint for solving the Q-IU problem. If we can improve LE, we can get a practical algorithm for the Q-IU problem that works much better than the literature for large k -values. (ii) We can see from Figure 2 that KL-LUCB increases faster as k increases. It is consistent with the theory that KL-LUCB depends on $k \log^2 k$ while (CB-)AL-Q-IU depends on $k \log k$. When k is large enough, (CB-)AL-Q-IU can perform better. (iii) In order sense, the performance of (CB-)AL-Q-IU is better than the literature. Thus, our work gives better theoretical insights about the Q-IU problem.

8 CONCLUSION

In this paper, we studied the problems of finding k top ρ fraction arms with an ϵ bounded error from a finite or infinite arm set. We considered both cases where the thresholds (i.e., λ_ρ and $\lambda_{[m]}$) are priorly known and unknown. We derived lower bounds on the sample complexity for all four settings, and proposed algorithms for them. For the Q-IK and Q-FK problems, our algorithms match the lower bounds. For the Q-IU and Q-FU problems, our algorithms are sample complexity optimal up to a log factor. Our simulations also confirm these improvements numerically.

Acknowledgements

This work has been supported in part by NSF ECCS-1818791, CCF-1758736, CNS-1758757, CNS-1446582, CNS-1314538, CNS-1717060; ONR N00014-17-1-2417; AFRL FA8750-18-1-0107.

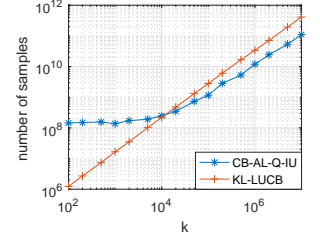


Figure 2: Comparison of CB-AL-Q-IU and KL-LUCB under prior \mathcal{F}_h . $\rho = 0.05$, $\epsilon = 0.1$, and $\delta = 0.01$.

References

- Agarwal, A., Agarwal, S., Assadi, S., and Khanna, S. (2017). Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*.
- Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*.
- Audibert, J. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *Conference on Learning Theory*.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*.
- Auer, P. and Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*.
- Aziz, M., Anderton, J., Kaufmann, E., and Aslam, J. (2018). Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*.
- Berry, D. A. and Eick, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in medicine*.
- Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). London: Chapman and Hall.
- Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*.
- Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*.
- Buccapatnam, S., Liu, F., Eryilmaz, A., and Shroff, N. B. (2017). Reward maximization under uncertainty: Leveraging side-observations on networks. *The Journal of Machine Learning Research*.
- Cao, W., Li, J., Tao, Y., and Li, Z. (2015). On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Advances in Neural Information Processing Systems*.
- Carpentier, A. and Valko, M. (2015). Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*.
- Chaudhuri, A. R. and Kalyan Krishnan, S. (2017). PAC identification of a bandit arm relative to a reward quantile. In *AAAI*.
- Chen, L., Gupta, A., and Li, J. (2016). Pure exploration of multi-armed bandit under matroid constraints. In *Conference on Learning Theory*.
- Even-Dar, E., Mannor, S., and Mansour, Y. (2002). PAC bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*. Springer.
- Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*.
- Goschin, S., Weinstein, A., Littman, M. L., and Chastain, E. (2013). Planning in reward-rich domains via PAC bandits. In *European Workshop on Reinforcement Learning*.
- Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*.
- Kalyan Krishnan, S. and Stone, P. (2010). Efficient selection of multiple bandit arms: Theory and practice. In *International Conference on Machine Learning*.
- Kalyan Krishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*.
- Kaufmann, E. and Kalyan Krishnan, S. (2013). Information complexity in bandit subset selection. In *Conference on Learning Theory*.
- Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. ACM.
- Liu, F., Wang, S., Buccapatnam, S., and Shroff, N. B. (2018). UCBoost: a boosting approach to tame complexity and optimality for stochastic bandits. In *International Joint Conference on Artificial Intelligence*. AAAI Press.
- Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*.
- Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*.