On Identifying Good Options under Combinatorially Structured Feedback in Finite Noisy Environments

Yifan Wu YWU12@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 CANADA

Abstract

We consider the problem of identifying a good option out of finite set of options under combinatorially structured, noisy feedback about the quality of the options in a sequential process: In each round, a subset of the options, from an available set of subsets, can be selected to receive noisy information about the quality of the options in the chosen subset. The goal is to identify the highest quality option, or a group of options of the highest quality, with a small error probability, while using the smallest number of measurements. The problem generalizes best-arm identification problems. By extending previous work, we design new algorithms that are shown to be able to exploit the combinatorial structure of the problem in a nontrivial fashion, while being unimprovable in special cases. The algorithms call a set multi-covering oracle, hence their performance and efficiency is strongly tied to whether the associated set multi-covering problem can be efficiently solved.

1. Introduction

Consider the problem of identifying the most rewarding option(s) out of finitely many. At your disposal are a number of probing devices, or just *probes*, that give you noisy measurements of the quality of a select set of options. More precisely, each *probe* is associated with a *known subset* of options whose quality the probe will measure. In a sequential process, the goal is to select the probes so that one can stop early to return, with high probability, a sufficiently rewarding option (or a set of options). As a specific example, consider the problem of identifying the segment on a road network that is in the worst shape after a long winter. Measurements can be obtained by sending trucks checking the road for potholes along the paths they travel on. The trucks must return to their garage every day. Here, the options correspond to road segments, the probes correspond to a closed walk in the road network that starts from the garage. Somewhat ironically, a road segment is "rewarding" (from the point of view of how beneficial it is to sending there the repair team) if it has many potholes.[1] Measurements are noisy, as potholes are easy to miss.

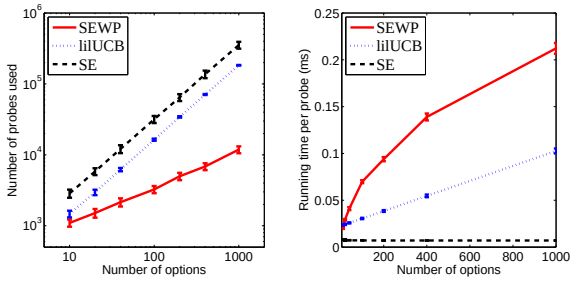

Figure 1. A specialized algorithm (SEWP) proposed in this paper can take nontrivial advantage of the probe structure as compared with simple adaptations of earlier algorithms, while being only marginally more expensive. All algorithms maintain the same error-rate. The plot on the left-hand-side uses a log-log-scale. Due to the special structure of the problem, the expected stopping time of the specialized algorithm scale linearly with $\sqrt{K}$, while the others scale linearly with K, the number of options.

Problems like the above one abound. Numerous quality assurance and surveying tasks are such that measurements give simultaneous information about multiple entities due to physical constraints on the measurement process. Application areas include technical computing (e.g., networking), biology (ecology, microbiology, etc.), physics, etc.

Of course, even though individual measurements might be

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

[1] In practice, one may want a whole "plan" at the end for the repair team. As often, we took the liberty of simplifying the problem to be able to focus on how the structure of probes should be used.

impossible, it is always possible to treat each probe as one that gives individual measurements for the options associated with it, though this could be wasteful (cf. Fig. 1). The main topic of the present paper is how to exploit, with efficient algorithms, when probes give information about multiple options.

The special case when each probe measures a single option, is known as the *best arm identification* problem, whose history goes back more than half a century (Bechhofer, 1958; Paulson, 1964), and with much activity in the last decade (e.g., Even-Dar et al. 2002, Mannor & Tsitsiklis 2004, Audibert et al. 2010, Kalyanakrishnan & Stone 2010, Bubeck et al. 2011, Kalyanakrishnan et al. 2012, Gabillon et al. 2012, Karnin et al. 2013, Kaufmann & Kalyanakrishnan 2013, Bubeck et al. 2013, Jamieson et al. 2014, Kaufmann et al. 2015, Zhou et al. 2014, Chen et al. 2014).

In this paper we consider two basic settings: identifying the best option with a prespecified error probability while using the smallest possible number of probes, and identifying a group of options of a fixed size, again with a prespecified error probability with the smallest possible number of probes. For the first setting, we propose two algorithms, SEWP and EGEWP described in Section 3, extending the works of Even-Dar et al. (2002) and Karnin et al. (2013). They work by constructing coverings with the probes of the sets of options not eliminated. The second algorithm removes a logarithmic term from the upper bound and it required a non-trivial extension of the median elimination method of Even-Dar et al. (2002). For the second setting, in Section 4, the quality of a group returned is assessed either by the quality of the worst option in the group (following Kalyanakrishnan & Stone (2010)), or by the average quality of options in the group (Zhou et al., 2014). We propose a single algorithm (SARWP) that essentially covers both cases. For the average quality, our distribution dependent upper bound is novel even in the bandit case and also near optimal in the worst case compared with the lower bound proposed by Zhou et al. (2014). For simple probe structures (singletons, or when a probe that covers all options is available), our algorithms are shown to be essentially unimprovable. We also give lower bounds for general probe structures. While both our lower and upper bounds express how the structure of the probes interferes with the structure of payoffs, they differ in subtle ways and it remains for future work to see whether there is a gap between them.

Due to space constraints, proofs and some experimental results are relegated to the appendix.

2. Preliminaries

In this section, we formulate the problem studied, as well as introducing the set covering problem, which will play an important role in our algorithms and analysis. We start by defining some notation.

2.1. Notation

The set of natural numbers will be denoted by $\mathbb{N}$, which includes zero. For a positive natural number n, $[n]$ denotes the set of integers between 1 and n: $[n] = \{1, \dots, n\}$. The power set, i.e., the set of all subsets of a set S, will be denoted by 2^S. As usual, functions, mapping set X to set Y will be viewed as elements of Y^X. For $v \in Y^X$, we will often write v_x instead of $v(x)$ to minimize clutter. This also helps with the next convention: When $U \subset X$, we will use v_U to denote the restriction of $v \in Y^X$ to U: $v_U(u) = v(u)$, $u \in U$. We identify $Y^{[n]}$ with Y^n (the set of n-tuples) in the natural way, which allows us to use notation v_U for $v \in Y^n \equiv Y^{[n]}$. The cardinality of a set S is denoted by $|S|$. Certain symbols will be reserved to denote elements of certain sets (i.e., p will always be an element of set $\mathcal{P}$). When using such reserved symbols, we will abbreviate (e.g.) $\sum_{p \in \mathcal{P}} f(p)$ to $\sum_p f(p)$. We will use $\log(\cdot)$ to denote the natural logarithm function.

2.2. Problem Formulation

A decision maker is given a pair $([K], \mathcal{P})$, where elements of $[K]$ are called arms, or, interchangeably, actions, and $\mathcal{P} \subset 2^{[K]}$ such that the sets in $\mathcal{P}$ cover $[K]$: $\cup \mathcal{P} = [K]$. Elements of $\mathcal{P}$ are called *probes*. A problem instance D, or *environment*, is specified by K distributions over the reals, $D = (D_1, \dots, D_K)$. The decision maker does not have direct access to these distributions. For $1 \le i \le K$, we think of distribution D_i as the distribution of "rewards" associated with arm i. We assume that the mean reward $\mu_i = \int x D_i(dx)$ of each arm is well defined. Further assumptions on D_i will be given later.

The goal of the decision maker is to find arms with the largest mean reward. For this, the decision maker can query the rewards of the arms by using the probes in a sequential manner. In particular, for each round $t = 1, 2, \dots$, first a random reward $X_{t,i} \sim D_i$ is generated for each arm i from its associated distribution. It is assumed that $X_{t,i}$ is independent of the other rewards $(X_{s,j})_{s \neq t \text{ or } j \neq i}$. We set $X_t = (X_{t,1}, \dots, X_{t,K}) \in \mathbb{R}^K$. In round $t = 1, 2, \dots$, the decision maker chooses a probe $p_t \in \mathcal{P}$ based on her past observations, to observe the values $X_{t,i}$ for each arm i in p_t; with our earlier introduced notation we can write that the decision maker observes $X_{t,p_t} \doteq (X_t)_{p_t} \in \mathbb{R}^{p_t}$. At the end of each round, the decision maker can decide between continuing or stopping to return a list of guesses (or a single guess) on the indices of the good arms. The goal is to stop as soon as possible, while avoiding poor guesses.

The following specific problem settings will be considered:

(i) *Fixed confidence, best-arm identification*. The *optimal arm* is unique: If $\mu^\star = \max_{i\in[K]} \mu_i$, $\max_{i:\mu_i\neq\mu^\star} \mu_i < \mu^\star$. The goal of the decision maker is to identify the index $i^\star = \operatorname{argmax}_{i\in[K]} \mu_i$ of the optimal arm. The decision maker is given a *confidence* parameter $0 \le \delta < 1$ and it is required that the guess returned after τ probes must be correct on an event $\mathcal{E}$ with probability at least $1-\delta$. Decision makers are compared based on their *probe complexity*, i.e., the number of probes they use when the "good event" $\mathcal{E}$ happens.

(ii) *PAC subset selection*. There are two subproblems that we consider. In both cases the decision maker is given a confidence, $0 \le \delta < 1$, a suboptimality threshold $\varepsilon > 0$ and a subset cardinality $1 \le m \le K$. The problems differ in how a quality $q(S,\mu)$ measure is assigned to a subset $S \subset [K]$ of arms. In both problems, the goal is to find a subset of arms of cardinality m such that $q(S,\mu) \ge \max_{P\subset[K]:|P|=m} q(P,\mu) - \varepsilon$ and with probability $1-\delta$, the decision maker must return a subset satisfying the above quality constraint. As before, decision makers are compared based on how many probes they use before stopping. The two quality measures considered are the reward of the worst arm in the set and the average reward: $q_{\min}(S,\mu) = \min_{i\in S} \mu_i$ and $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i\in S} \mu_i$, $S \subset [K]$, $|S| = m$. We call the corresponding problems the *strong* and the *average* PAC subset selection problems.

An algorithm used by a decision maker to select probes, stop and return a guess will be said to be *admissible* with respect to a class of environments, if, for *any* environment within the class and any $0 \le \delta < 1$, the guess computed is correct (according to the previous requirements) with probability $1-\delta$.

The above problems have been considered in the past in the special case when $\mathcal{P}$ contains singletons only, by a number of authors (see Section 1 for some references). We shall call these the "bandit" problems. While one can readily apply the algorithms developed for the bandit case to our problem, the expectation is that the probe complexity of reasonable algorithms should improve considerably as $\mathcal{P}$ becomes "richer" (this was illustrated in Fig. 1). The question is how the structure of $\mathcal{P}$ together with the problem instance influences the problem complexity. For example, in the extreme case when $\mathcal{P}$ contains $[K]$, we expect the probe complexity of reasonable algorithms to scale sublinearly with K, whereas in the bandit case a linear scaling is unavoidable. The case when $\mathcal{P} = \{[K]\}$ will be called the *full information case*.

Note that since all probes "cost" the same amount (one unit of time), a reasonable algorithm will avoid any probe p that is entirely included in some other probe $p' \in \mathcal{P}$. Hence, we may as well assume that the set of probes does not have nontrivial chains in it.

We will present results for the class of environments $\mathcal{D}_{\text{sg}}$ with the following restrictions: For each $1 \le i \le K$, D_i is sub-Gaussian with common parameter $\sigma^2 = 1/4$:

$$\log \int_{\mathbb{R}} e^{-\lambda(x-\mu_i)} D_i(dx) \le \lambda^2\sigma^2/2 = \lambda^2/8$$

for all $\lambda \in \mathbb{R}$. To simplify the presentation of our results, without loss of generality, we *assume that* $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_K$. (note that, obviously, the algorithms do not use this assumption). For further simplicity, we assume that $\Delta_i \in [0,1]$ for all $i \in [K]$ where $\Delta_i = \mu_1 - \mu_i$, $2 \le i \le K$. Our assumptions on the reward distributions D_i are satisfied if, for example, D_i has bounded support.

We will present algorithms, which will be shown to be admissible for $\mathcal{D}_{\text{sg}}$ and we will bound their probe complexities. The bounds on the probe complexities will be given in terms of the *(suboptimality) gaps* Δ_i, $2 \le i \le K$, i.e., they will be dependent on the distributions $D = (D_1, \ldots, D_K)$. Hence, we call them distribution dependent bounds. We will accompany our constructive results with lower bounds, putting a lower limit on the probe complexity of all admissible algorithms. Again, these will be given in terms of the gaps Δ_i.

2.3. Set Multi-Cover Problems

Probes allow one to "explore" multiple arms simultaneously. Clever algorithms should use the probes in a smart way to guarantee the necessary number of samples for each of the arms while using the smallest number of probes. If, for example, $n \in \mathbb{N}$ observations are enough from each of the arms to distinguish their mean payoff from that of the optimal arm, then an intelligent algorithm would try to create the smallest *covering* of $[K]$ using the subsets in $\mathcal{P}$ to meet this requirement. More generally, for $J \subset [K]$, we define

$$\mathcal{C}_{\text{IP}}(J,n) = \min\left\{ \textstyle\sum_p s_p : s \in \mathbb{N}^{\mathcal{P}}, \sum_{p:i\in p} s_p \ge n, i \in J \right\}$$

to be the *cost* of the smallest n-fold multi-covering of elements of J. Any $s \in \mathbb{N}^{\mathcal{P}}$ achieving the minimum is called an *optimal (integral) n-cover* of J, while a feasible vector s is called an n-cover. Given an n-cover $s \in \mathbb{N}^{\mathcal{P}}$, we will say that probe p belongs to s (writing $p \in s$) if $s_p > 0$. The optimization problem defining $\mathcal{C}_{\text{IP}}$ is a linear integer program (hence the IP in $\mathcal{C}_{\text{IP}}$). Relaxing the *integrality* constraint $s \in \mathbb{N}^{\mathcal{P}}$ to the nonnegativity constraint $s \in [0,\infty)^{\mathcal{P}}$, we get a so-called *fractional optimal n-cover* of J by solving the otherwise identical optimization problem. The resulting optimal value will be denoted by $\mathcal{C}_{\text{LP}}(J,n)$. Note that

the relaxed problem is a linear program, explaining "LP" in $\mathcal{C}_{\mathrm{LP}}$. While this linear program has potentially exponentially many variables in K, it can still be efficiently solved provided an efficiently computable membership oracle is available for its dual (Grötschel et al., 1993). Both $\mathcal{C}_{\mathrm{IP}}(J,n)$ and $\mathcal{C}_{\mathrm{LP}}(J,n)$ can be extended to non-integer values of n.

It follows immediately from the definitions that $\mathcal{C}_{\mathrm{LP}}(J,n) \le \mathcal{C}_{\mathrm{IP}}(J,n)$. Further, for any $a > 0$, $\mathcal{C}_{\mathrm{LP}}(J,a\,n) = a\,\mathcal{C}_{\mathrm{LP}}(J,n) = an\,\mathcal{C}_{\mathrm{LP}}(J,1)$. The *integrality gap* for a set multi-covering problem instance is given by $(\mathcal{P},J,n)$ is $\mathcal{C}_{\mathrm{IP}}(J,n)/\mathcal{C}_{\mathrm{LP}}(J,n)$ (Vazirani, 2001).

Our algorithms will need "small" n-covers for various subsets $J \subset [K]$. Depending on the structure of $\mathcal{P}$, calculating an optimal multi-cover of J may be easy or hard[2] (e.g., Slavik, 1998; Schrijver, 2003; Korte & Vygen, 2006). Thus, to keep the presentation general, our algorithms will rely on a set multi-covering *oracle* COrcl, which given $J,n,\mathcal{P}$, returns an n-fold multi-cover of J using the sets in $\mathcal{P}$. Denote by $\mathcal{C}_{\mathrm{O}}(J,n)$ the cost of the multi-cover returned by the oracle on J,n (as with $\mathcal{C}_{\mathrm{IP}}$ and $\mathcal{C}_{\mathrm{LP}}$ the dependence on $\mathcal{P}$ is suppressed). The oracle's integral (fractional) approximation gap, $\mathcal{G}_{IP}(O,\mathcal{P})$ ($\mathcal{G}_{LP}(O,\mathcal{P})$), is the worst-case multiplicative loss due to using COrcl in place of an optimal integral (fractional) cover. In particular, with $\star \in \{IP, LP\}$,

$$\mathcal{G}_\star(O,\mathcal{P}) = \sup_{n\in\mathbb{N}^+, J\subset[K]} \frac{\mathcal{C}_{\mathrm{O}}(J,n)}{\mathcal{C}_\star(J,n)}\,.$$

Let $d = \max_{p\in\mathcal{P}} |p|$ be the maximum number of actions that can be covered by a single probe. If the set-system $\mathcal{P}$ has no special structure, one possibility is to use the greedy algorithm G as the oracle. This algorithm works by sequentially setting $s_p = n$ for the probe $p \in \mathcal{P}$ that covers the maximum number of active arms in J and then deactivates the arms that are covered by p, until all arms are deactivated. Further, $\mathcal{G}_{LP}(O,\mathcal{P}) \le 1 + \log(d) \le 1 + \log(K)$. Lovász (1975) showed that $\mathcal{C}_{\mathrm{G}}(J,1) \le (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,1)$. Then, $\mathcal{C}_{\mathrm{G}}(J,n) = n\,\mathcal{C}_{\mathrm{G}}(J,1) \le (1+\log d)n\,\mathcal{C}_{\mathrm{LP}}(J,1) = (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,n)$, showing that the required inequality indeed holds. Raz & Safra (1997) proved that the exists some constant $c > 0$ such that, unless $P = NP$, no approximation ratio of $c\log(K)$ can be achieved, so in a worst-case the greedy algorithm is a near-optimal approximation algorithm.

3. Finding the Best Arm

In this section we present two algorithms and their analysis for the fixed confidence, best-arm identification problem. Recall that in this problem, given a set of probes $\mathcal{P}$ and a confidence $\delta \in (0,1]$, we need to design a sequential procedure that identifies the best arm $i^\star$ with probability at least $1-\delta$ using as few probes as possible.

3.1. Successive Elimination with Probes

The first algorithm modifies the successive elimination algorithm of Even-Dar et al. (2002) to take into account the richer observation structure of our problem. Recall that the algorithm of Even-Dar et al. (2002) works in phases, in each phase observing a certain number of rewards for each remaining candidate actions. At the end of the phase the provably suboptimal actions are eliminated. The number of observations in each phase depends only on the phase index. The process stops when the candidate set contains a single element. The main difference to the algorithm of Even-Dar et al. (2002) is that in each phase our algorithm, which we call Successive Elimination with Probes (SEWP), computes a set multi-covering for the remaining candidate actions given the probes, with a requirement adjusted to the phase index. The returned multi-cover is then used to get the observations for the remaining actions.

Algorithm 1 SuccessiveEliminationWithProbes (SEWP)

Inputs: K, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0,1] \to [0,\infty)$.
Initialize candidate set: $A_1 = [K]$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each p in $C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for arm i.
  $A_{t+1} \leftarrow \{i \in A_t : \hat{\mu}_i(t) + 2g(t,\delta) > \max_{j\in A_t} \hat{\mu}_j(t)\}$.
  if $|A_{t+1}| = 1$ **then**
    Return the arm in A_{t+1}.
  end if
end for

Our first result shows that Algorithm 1 is admissible and gives an upper bound on its probe complexity. To state it, define the scheduling and confidence functions

$$f(t) = 2^t, \qquad g(t,\delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}\,. \tag{1}$$

For simplicity, assume that the arms are ordered in decreasing order of their mean rewards and $\Delta_2 > 0$, i.e., the optimal arm is unique. For $2 \le i \le K$ define

$$\widehat{T}_i(\delta) = 1 + \max\left\{ s \,:\, g(s,\delta) \ge \frac{\Delta_i}{4} \right\}, \tag{2}$$

$$\hat{N}_i(\delta) = \frac{128}{\Delta_i^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_i} \right) \tag{3}$$

[2] Computing the exact solution for the decision version of set covering (i.e., when $n = 1$), when $\mathcal{P}$ can be any covering system, is known to be NP-hard (Vazirani, 2001).

and let $\widehat{T}_{K+1}(\delta) = 0$ and $\hat{N}_{K+1}(\delta) = 0$. Note that $2^{\widehat{T}_i(\delta)+1} \leq \hat{N}_i(\delta)$, and both are decreasing with $i \geq 2$ increasing.

Theorem 1. *Pick any $0 \leq \delta < 1$ and let SEWP run with inputs $(K, \delta, \mathcal{P}, f, g)$ with f, g given by (1). Then, with probability at least $1-\delta$, SEWP returns the optimal arm $i^\star = 1$ within N probes, where N satisfies*

$$N \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}(\delta)+1}^{\widehat{T}_i(\delta)} \mathcal{C}_{\mathrm{IP}}([i], 2^t) . \quad (4)$$

Furthermore, with $\hat{M}_i(\delta) \doteq \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta)$,

$$N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_i(\delta)\, \mathcal{C}_{\mathrm{LP}}([i], 1) . \quad (5)$$

The bound (4) may be tighter than that shown in (5), but perhaps the second is a bit easier to understand.[3] For simplicity, let us explain (5). Once (5) is explained, the meaning of (4) follows. The term $\mathcal{G}_{LP}(O, \mathcal{P})$ is the price of using an oracle combined with some upper bounding that allowed us to arrive at this simpler result by resorting to the linearity properties of $\mathcal{C}_{\mathrm{LP}}$. The rest is what we call a sequential fractional multi-cover with the requirements that arm i be covered $\hat{N}_i(\delta)$ times: In a sequential multi-cover, the covering is not done in a single-shot, but is done in phases. In the first phase, all the arms must be covered $\hat{M}_K(\delta)$ times. In the next phase, all the arms but the last must be covered $\hat{M}_{K-1}(\delta)$ times, etc., up to the last phase when arms one and two must be covered $\hat{M}_2(\delta)$ times. Note that the total requirements for an arm i are $\hat{M}_K(\delta) + \hat{M}_{K-1}(\delta) + \cdots + \hat{M}_i(\delta) = \hat{N}_K(\delta) - \hat{N}_{K+1}(\delta) + \hat{N}_{K-1}(\delta) - \hat{N}_K(\delta) + \cdots + \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta) = \hat{N}_i(\delta)$. Roughly $\hat{N}_i(\delta) \approx O(1/\Delta_i^2)$ is the number of observations needed from arm i (and one) in order to be able to tell which of the two arms has a bigger mean reward. Now, compared to (5), (4) uses a more precise expression for the number of probes, by relying on the the phase structure of the algorithm.

An alternative choice of $f(t)$ and $g(t, \delta)$ is that $f(t) = 1$ and $g(t, \delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{t}}$, which leads to $\hat{N}_i(\delta) = O\left(\frac{1}{\Delta_i^2} \log \frac{K}{\delta \Delta_i}\right)$ instead.

The proof, which borrows ideas from Even-Dar et al. (2002), is in Appendix A.1. To prove that SEWP is admissible, one only needs to show that when none of the confidence intervals based on g used in the elimination step fail, the optimal arm will not be eliminated. This essentially relied on Hoeffding's inequality, union bounds and calculations. To calculate the bound on the probe complexity bound, one shows that arm i will be eliminated after phase $\hat{T}_i(\delta)$. This happens because in each phase the confidence sets of all arms decrease at a uniform rate.

Now, we argue that this bound is tight up to a $\log K$ factor, at least in some cases. In particular, in the bandit case, the covering problem is trivial and we can use an optimal covering oracle. Then, $\mathcal{C}_{\mathrm{O}}([i], 2^t) = i2^t$, and hence the bound becomes $O\left(\sum_{i=1}^{K} \frac{1}{\Delta_i^2} \log\left(\frac{K}{\delta} \log \frac{1}{\Delta_i}\right)\right)$. Up to a log factor, this matches the lower bound of Kaufmann et al. (2015) which takes the form $\Omega\left(\sum_{i=1}^{K} \Delta_i^{-2} \log(1/\delta)\right)$. Furthermore, as noted by Jamieson et al. (2014) (based on a result of Farrell (1964)) the $\log\log \Delta^{-1}$ term is necessary.

To examine the tightness of the upper bound, we derive a distribution dependent lower bound on the probe complexity of algorithms admissible for $\mathcal{D}_{\mathrm{sg}}$. Call an environment D a Gaussian environment with common variance σ^2 if for any $1 \leq i \leq K$, D_i is a Gaussian with variance σ^2.

Theorem 2 (Distribution-dependent lower bound)**.** *For any algorithm admissible for $\mathcal{D}_{sg}$, any confidence $0 < \delta < 1/2$, any probe set $\mathcal{P}$, any sequence $0 = \Delta_1 < \Delta_2 \leq \ldots \Delta_K$, if D is a Gaussian environment with common variance $\sigma^2 = 1/4$ and means $\mu_1 = \mu_2 + \Delta_2 = \cdots = \mu_K + \Delta_K$, if N is the number of probes used by the algorithm on D then*

$$\mathbb{E}[N] \geq \min_{s \in [0,\infty)^{\mathcal{P}}} \sum_{p \in \mathcal{P}} s_p \quad \textit{s.t.} \sum_{p: 1 \in p} s_p \geq \frac{1}{4\Delta_2^2} \log \frac{1}{6\delta} ,$$
$$\textit{and} \sum_{p: i \in p} s_p \geq \frac{1}{4\Delta_i^2} \log \frac{1}{6\delta} , \quad 2 \leq i \leq K .$$

The proof can be found in Appendix A.2.

Note that the lower bound clearly reflects the structure of $\mathcal{P}$. However, even disregarding the constants and logarithmic factors, there is still a gap between our upper and lower bounds: In the upper bound, as explained before, the size of a *sequential* cover that appears, while in the lower bound, the size of a "one-shot" cover is seen. Note that in either the bandit or the full information case, there is no gap between these quantities. We were able to establish a gap of $\log(K)$ when considering sequential and one-shot *integral* covers. However, it remains a very interesting open question whether the gap can be closed in the fractional case.

3.2. An Alternative Algorithm to Find the Best Arm

The second algorithm is a generalization of the exponential gap elimination algorithm of Karnin et al. (2013), which improves the logarithmic term in the sample complexity from $\log(\frac{K}{\delta} \log \frac{1}{\Delta})$ to $\log(\frac{1}{\delta} \log \frac{1}{\Delta})$ for the bandit problem. So we expect that generalizing that algorithm to our setting will have a similar improvement regarding the $\log K$ term.

The exponential gap elimination algorithm of Karnin et al.

[3] In fact, if $\mathcal{C}_{\mathrm{O}}(\cdot, n)$ is monotone increasing, (4) will hold with $\mathcal{C}_{\mathrm{O}}$ replacing $\mathcal{G}_{IP} \cdot \mathcal{C}_{\mathrm{IP}}$, further tightening the bound.

(2013) calls the median elimination algorithm of Even-Dar et al. (2002) as a subroutine, which finds an ε-optimal arm using $O(K\varepsilon^{-2}\log(1/\delta))$ samples with probability at least $1-\delta$ (an arm is ε-optimal iff its expected reward is at least $\mu_1-\varepsilon$). So before generalizing the exponential gap elimination algorithm, we need to first design a counterpart for the median elimination algorithm.

3.2.1. Median Elimination With Probes

Simply replacing the uniform sampling in each phase in the median elimination algorithm of Even-Dar et al. (2002) with a set multi-cover does not work (shown in Appendix B.1), so a more careful design is needed. Our proposed algorithm, called Median Elimination With Probes (MEWP) is shown in Algorithm 2. It essentially runs the original median elimination algorithm for bandits over a one-cover of all arms (that is, each probe in the cover is treated as an arm in the bandit setting), and in each phase we eliminate half of the *probes* that do not seem to cover a good arm. We stop running median elimination when a single probe covers all the remaining arms. Then the algorithm enters its second stage where we use this probe until we identify an almost optimal arm from the remaining ones. In the next theorem we prove that the algorithm is admissible, and give an upper bound on the number of probes required to find an ε-optimal arm.

Algorithm 2 MedianEliminationWithProbes

Inputs: K, $\delta\in(0,1]$, $\varepsilon>0$, $\mathcal{P}$.
Set $\varepsilon_t=\frac{\varepsilon}{6}(\frac{3}{4})^t$, $\delta_t=\frac{\delta}{2^{t+1}}$.
$C\leftarrow$ COrcl$([K],1,\mathcal{P})$, and define a partition of the arms as $A_1=\{\pi_p\subset p\,:\,p\in C,\cup_{p\in C}\pi_p=[K]\}$.
for $t=1,2,\ldots$ **do**
  for all $\pi\in A_t$ **do**
    Use $\frac{4}{\varepsilon_t^2}\log\frac{3|\pi|}{\delta_t}$-times $p\in C$ that covers π to get observations for each arm in p.
    Let $\hat{\mu}_\pi(t)=\max_{i\in\pi}\hat{\mu}_i(t)$, where $\hat{\mu}_i(t)$ is the empirical mean reward of arm i based on the observations in the actual phase t.
  end for
  Find the median $m(t)$ of $\{\hat{\mu}_\pi(t):\pi\in A_t\}$.
  Let $A_{t+1}=\{\pi\in A_t\,:\,\hat{\mu}_\pi(t)\geq m(t)\}$.
  if $|A_{t+1}|=1$ **then**
    terminate the loop and let $\widehat{\pi}^*$ be the single element of A_{t+1}
  end if
end for
If $|\widehat{\pi}^*|>1$, use the probe that covers $\widehat{\pi}^*$ for $\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$-times.
Return the arm $\hat{i}^*\in\widehat{\pi}^*$ with the highest empirical mean based on these observations.

Theorem 3. *With probability at least $1-\delta$, MEWP returns an ε-optimal arm $\hat{i}^*$, and N, the total number of probes used by the algorithm is*

$$N=O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right). \tag{6}$$

where $|\pi_{\max}|=\max_{\pi\in A_1}|\pi|$.

Note that we have $|\pi_{\max}|$ inside the log term instead of the expected 1. It can be shown that the argument of the log term cannot be 1 in our problem setting, at least in the full information case (where it has to be K). Detailed discussion about this can be found in Appendix B.2.

3.2.2. Exponential Gap Elimination Algorithm

Algorithm 3 ExpGapEliminationWithProbes

Inputs: K, δ, $\mathcal{P}$.
Initialize candidate set: $A_1=[K]$. Set $\varepsilon_t=\frac{1}{4\cdot 2^t}$, $\delta_t=\frac{\delta}{50t^3}$.
for $t=1,2,\ldots$ **do**
  $C(t)\leftarrow$ COrcl$(A_t,1,\mathcal{P})$.
  Create a partition Π_t of A_t such that $\Pi_t=\{\pi_p\subset p\,:\,p\in C(t),\cup_{p\in C(t)}\pi_p=A_t\}$.
  for $\pi_p\in\Pi_t$ **do**
    Use probe p for $\frac{2}{\varepsilon_t^2}\log\frac{2|\pi_p|}{\delta_t}$-times to get observations for each arm in p.
  end for
  For each $i\in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations in phase t for arm i.
  $i_t\leftarrow$ MedianEliminationWithProbes$(A_t,\frac{\varepsilon_t}{2},\delta_t)$.
  Let $A_{t+1}=\{i\in A_t\,:\,\hat{\mu}_i(t)\geq\hat{\mu}_{i_t}(t)-\varepsilon_t\}$.
  if $|A_{t+1}|=1$ **then**
    Return the arm in A_{t+1}.
  end if
end for

Given the MEWP algorithm, we continue with generalizing the exponential gap elimination algorithm. The new algorithm, called Exponential Gap Elimination with Probes (EGEWP), is shown in Algorithm 3. The new idea here is to use the partition-based exploration technique (as in the MEWP algorithm) and replace the bandit-case median elimination subroutine with MEWP. The analysis follows a combination of the techniques of Karnin et al. (2013) and the proof of Theorem 3. However, due to the more complicated observation structure, we are only able to prove a Δ_2 dependent upper bound on the number of probes:

Theorem 4. *If the oracle COrcl always returns the optimal solution for integer programming, EGEWP finds the optimal arm with probability at least $1-\delta$ after using*

$$O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right) \tag{7}$$

probes where $|p_{\max}|=\max_{p\in\mathcal{P}}|p|$.

If COrcl is not guaranteed to return the optimal integer cover, the above theorem still holds by making the following modification to the algorithm to ensure that Π_{t+1} is not worse than Π_t for every t: if $|\{\pi \in \Pi_t : \pi \cap A_{t+1} \neq \varnothing\}| < \mathcal{C}_{\mathrm{O}}(A_{t+1}, 1)$, then use the same partition pattern from Π_t for Π_{t+1}.

Compared to the bound for SEWP, the $\log K$ term is replaced with $\log |p_{\max}|$. More specifically, in the full information case, the upper bound becomes $O\left(\frac{1}{\Delta_2^2} \log\left(\frac{K}{\delta} \log \frac{1}{\Delta_2}\right)\right)$, which is the same as the upper bound for SEWP. In the bandit case, the algorithm is exactly the same as the exponential gap elimination algorithm of Karnin et al. (2013), which enjoys an optimal $O\left(\sum_{i=1}^{K} \frac{1}{\Delta_i^2} \log\left(\frac{1}{\delta} \log \frac{1}{\Delta_i}\right)\right)$ upper bound on the number of probes, and is better than the upper bound for SEWP in bandit case. Therefore, although not formally proved, we expect that EGEWP enjoys an improved probe complexity compared with SEWP.

4. PAC Subset Selection

In this section, we consider the two PAC subset selection problems introduced in Section 2. The first, named *strong* PAC subset selection, is the same as the EXPLORE-m problem introduced by Kalyanakrishnan & Stone (2010) where the goal is to find m (ε, m)-optimal arms. The second problem, named *average* PAC subset selection, is to select a subset of m arms with ε-optimal average reward, introduced by Zhou et al. (2014).

The basic idea of our approach is to generalize our SEWP algorithm with two modifications: (i) First, besides rejecting the arms that cannot be in the best m arms after each phase, we also accept arms that have enough confidence to be one of the best m arms, which shares a similar idea with the Racing algorithm in Kaufmann & Kalyanakrishnan (2013). (ii) Specific stopping conditions are designed to meet the ε-relaxation in the problem definition.

To make it easier to express the probe complexity, we introduce a new symbol $\Delta_i^{(\varepsilon,m)}$ defined by $\Delta_i^{(\varepsilon,m)} = \max\{\mu_i - \mu_{m+1}, \varepsilon\}$ if $i \leq m$ and $\Delta_i^{(\varepsilon,m)} = \max\{\mu_m - \mu_i, \varepsilon\}$ if $i > m$. We then sort $\Delta_i^{(\varepsilon,m)}$ for all $i \in [K]$ in ascending order and let $S_{(i)}$ be the first i arms in the list, while $\Delta_{(i)}^{(\varepsilon,m)}$ denotes the i-th smallest entry.

Analogously to Theorem 1, let $f(t) = 2^t$, $g(t,\delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}$, and define

$$\hat{N}_{(i)}(\varepsilon,\delta) = \frac{128}{\left(\Delta_{(i)}^{(\varepsilon,m)}\right)^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_{(i)}^{(\varepsilon,m)}}\right) \quad (8)$$

and let $\hat{N}_{(K+1)}(\varepsilon,\delta) = 0$.

Note that $\hat{N}_{(1)}(\varepsilon,\delta) = \hat{N}_{(2)}(\varepsilon,\delta)$ since $\Delta_{(1)}^{(\varepsilon,m)} = \Delta_{(2)}^{(\varepsilon,m)} = \max\{\mu_m - \mu_{m+1}, \varepsilon\}$. Also let $\hat{M}_{(i)}(\varepsilon,\delta) \doteq \hat{N}_{(i)}(\varepsilon,\delta) - \hat{N}_{(i+1)}(\varepsilon,\delta)$.

4.1. *Strong* PAC Subset Selection

First we propose an algorithm that returns a subset $\hat{S}^*$ containing m (ε, m)-optimal arms with high probability. An arm i is defined to be (ε, m)-optimal iff $\mu_i \geq \mu_m - \varepsilon$. This requirement is the same as $q_{\min}(\hat{S}^*, \mu) \geq q_{\min}([m], \mu) - \varepsilon$ where $q_{\min}(S, \mu) = \min_{i \in S} \mu_i$.

The algorithm, called Successive Accept Reject with Probes (SARWP) is shown in Algorithm 4. The following theorem shows that Algorithm 4 is admissible and the probe complexity is bounded.

Algorithm 4 SuccessiveAcceptRejectWithProbes

Inputs: K, m, ε, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0,1] \to [0, \infty)$.
Initialize candidate set $A_1 = [K]$, accepted arms $A_1^a = \varnothing$, rejected arms $A_1^r = \varnothing$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each $p \in C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for arm i. Sort the arms in A_t in descending order of $\hat{\mu}_i(t)$. Let H_t be the first $m - |A_t^a|$ arms and $L_t = A_t \setminus H_t$.
  if $\min_{i \in H_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) + 2g(t,\delta) - \varepsilon$ **then**
    Return $\hat{S}^* = A_t^a \cup H_t$ as selected subset.
  end if
  Let
  $A_{t+1}^a = A_t^a \cup \{i \in H_t : \hat{\mu}_i(t) > \max_{j \in L_t} \hat{\mu}_j(t) + 2g(t,\delta)\}$,
  $A_{t+1}^r = A_t^r \cup \{i \in L_t : \hat{\mu}_i(t) < \min_{j \in H_t} \hat{\mu}_j(t) - 2g(t,\delta)\}$,
  and $A_{t+1} = [K] - A_{t+1}^a - A_{t+1}^r$
end for

Theorem 5. *With probability at least $1 - \delta$, SARWP returns a subset $\hat{S}^*$ of size m within N probes, where $q_{\min}(\hat{S}^*, \mu) \geq q_{\min}([m], \mu) - \varepsilon$ and N satisfies $N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_{(i)}(\varepsilon,\delta) \mathcal{C}_{\mathrm{LP}}\left(S_{(i)}, 1\right)$.*

The upper bound on the probe complexity is in a similar form to the one for SEWP in Theorem 1, while here the number of samples required for arm i is determined by $\Delta_i^{(\varepsilon,m)}$ instead of Δ_i. This complexity measure matches existing work for the bandit case (Kalyanakrishnan et al.,

2012; Kaufmann & Kalyanakrishnan, 2013). In the bandit case, the upper bound matches the worst case lower bound in Kalyanakrishnan et al. (2012): $\Omega(K\varepsilon^{-2}\log(m/\delta))$, up to logarithmic factors. We do not have a distribution dependent lower bound like Theorem 2 and even in the bandit case a distribution dependent lower bound for $\varepsilon > 0$ is still an open question (Kaufmann & Kalyanakrishnan, 2013).

4.2. *Average* PAC Subset Selection

Next we consider the problem that aims to find a subset whose aggregate regret is ε-optimal. Given a subset $S \subset [K]$ and $|S| = m$, the *aggregate regret* of S is defined as $R_S = \frac{1}{m}\left(\sum_{i\in[m]}\mu_i - \sum_{i\in S}\mu_i\right) = q_{\text{avg}}([m],\mu) - q_{\text{avg}}(S,\mu)$ where $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i\in S}\mu_i$. The aggregate regret of S is said to be ε-optimal iff $R_S \le \varepsilon$.

To address the problem of finding an average ε-optimal subset, Algorithm 4 can still be employed by only modifying the stopping condition according to the different objective. The new stopping condition is described as follows:

Stopping condition for average PAC subset selection: First for each $i \in A_t$, we construct an adversarial estimation $\hat{\mu}'_i(t)$ by setting $\hat{\mu}'_i(t) = \hat{\mu}_i(t) - g(t,\delta)$ if $i \in H_t$ and $\hat{\mu}'_i(t) = \hat{\mu}_i(t) + g(t,\delta)$ if $i \in L_t$. Then we sort the arms in descending order according to $\hat{\mu}'_i(t)$ and let H'_t be the first $m - |A^a_t|$ arms while L'_t be the remaining. The algorithm stops and returns $\hat{S}^* = A^a_t \cup H_t$ if

$$\sum_{i\in H_t\setminus H'_t}(\hat{\mu}_i(t) - g(t,\delta)) \ge \sum_{i\in H'_t\setminus H_t}(\hat{\mu}_i(t) + g(t,\delta)) - m\varepsilon\,.$$

This way of constructing "adversarial estimation" is similar to the one in the CLUCB algorithm of Chen et al. (2014), where the goal is to identify a subset with the highest reward sum without ε relaxation.

The next theorem shows that with the modified stopping condition, Algorithm 4 is admissible and bounds its probe complexity. Define

$$b(m,\varepsilon) = \max\left\{a \in \mathbb{N}^+ : \mu_{m-a+1} - \mu_{m+a} \le \frac{m\varepsilon}{a}\right\}, \tag{9}$$

or $b(m,\varepsilon) = 1$ if $\mu_m - \mu_{m+1} > m\varepsilon$. Then we have the following result:

Theorem 6. *With probability at least $1-\delta$, SARWP with modified stopping condition returns a subset $\hat{S}^*$ of size m within N probes, where $q_{avg}(\hat{S}^*,\mu) \ge q_{avg}([m],\mu) - \varepsilon$ and N satisfies $N \le \mathcal{G}_{LP}(O,\mathcal{P})\sum_{i=2}^{K}\hat{M}_{(i)}(m\varepsilon/b,\delta)\mathcal{C}_{\text{LP}}\left(S_{(i)},1\right)$, where $b = b(m,\varepsilon)$.*

Compared with Theorem 5, the complexity here is measured by $\Delta_i^{(m\varepsilon/b,m)}$ instead. This distribution dependent complexity measure is novel even in the bandit case since the algorithm in Zhou et al. (2014) comes with distribution independent guarantee only. Regarding the worst case performance, since $b(m,\varepsilon) \le \min\{m, K-m\}$, in bandit case our upper bound can be further relaxed to $O\left(\frac{K}{\varepsilon^2}\log\left(\frac{K}{\delta}\log\frac{1}{\varepsilon}\right)\right)$ if $m \le K/2$ and $O\left(\frac{K(K-m)^2}{m^2\varepsilon^2}\log\left(\frac{K}{\delta}\log\frac{K-m}{m\varepsilon}\right)\right)$ if $m > K/2$. Compared with the worst case lower bound in Zhou et al. (2014): $\Omega\left(\frac{K}{\varepsilon^2}\left(1+\frac{\log(1/\delta)}{m}\right)\right)$ for $m \le K/2$ and $\Omega\left(\frac{K-m}{m}\cdot\frac{K}{\varepsilon^2}\left(\frac{K-m}{m}+\frac{\log(1/\delta)}{m}\right)\right)$ for $m > K/2$, although our upper bound does not exactly match this worse case lower bound, our distribution dependent quantity $b(m,\varepsilon)$ shows how the different $\frac{K}{\varepsilon^2}$ and $\frac{K(K-m)^2}{m^2\varepsilon^2}$ terms appear for small m and large m compared with $K/2$.

5. Conclusions

We introduced a generalized version of the best arm identification problem, where a decision maker can query multiple arms at a time. This generalization describes several real world problems that are not adequately modeled by the standard best-arm identification problem. We generalized several existing algorithms and provided distribution dependent upper and lower bounds on the probe complexity, and showed that our algorithms achieve essentially the best possible performance in special cases. In the PAC subset selection problems our complexity measure either matches existing works for the bandit case or provides some new insights. One very interesting question that remains for future work is whether there is a real gap between our lower and upper bounds. However, much work remains to be done: We view our paper as opening a whole new practical and exciting research area of investigating richer feedback structures in "winner selection" problems. Interesting questions include how to change the algorithms when the subsets to be returned are restricted, or when probes are associated with costs.

Acknowledgement

This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning (AICML) and NSERC.

References

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

Bechhofer, R. E. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pp. 255–270, 2002.

Farrell, R. H. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, 35(1):36–72, 1964.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Grötschel, M., Lovász, L., and Schrijver, A. *Geometric Algorithms and Combinatorial Optimization*. Springer, 2 edition, 1993.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2014.

Kalyanakrishnan, S. and Stone, P. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of International Conference on Machine Learning (ICML)*, 2010.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

Kaufmann, E. and Kalyanakrishnan, S. Information complexity in bandit subset selection. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2013.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015. (to appear).

Korte, B. H. and Vygen, J. *Combinatorial optimization: theory and algorithms*. Springer, 3 edition, 2006.

Lovász, L. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 13(4):383–390, 1975.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

Paulson, E. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

Raz, R. and Safra, S. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *STOC*, pp. 475–484, 1997.

Schrijver, Alexander. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2003.

Slavik, Petr. *Approximation Algorithms for Set Cover and Related Problems*. PhD thesis, State University of New York at Buffalo, 1998. AAI9833643.

Vazirani, V.V. *Approximation algorithms*. Springer, 2001.

Zhou, Y., Chen, X., and Li, J. Optimal pac multiple arm identification with applications to crowdsourcing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2014.

A. Proofs

A.1. Proof of Theorem 1

We start with a technical lemma:

Lemma 7. *Let $0 < a < 1/e$, $b \geq 2$. Then, for any $n \geq n^*(a,b) \doteq \frac{2}{a}\log(2b\log\frac{1}{a})$, $an \geq \log(b\log n)$.*

Proof. Let $q_1(x) = ax$, $q_2(x) = \log(b\log x) = \log(\log x) + \log b$, $x > e$. The claim to be proven is that for any $n \geq n^* \doteq n^*(a,b)$, $q_1(n) \geq q_2(n)$. By differentiation, it is easy to verify that the function $f(x) = q_1(x) - q_2(x)$ is non-decreasing if and only if $x\log x \geq 1/a$. Hence, it suffices to show that $n^*\log n^* \geq 1/a$, $n^* > e$ so that $q_2(n^*)$ is well-defined and $q_1(n^*) \geq q_2(n^*)$.

From the assumptions and the definition of n^*, we get that $n^* \geq 2\log(4)/a \geq \frac{1}{a} > e$. Hence $q_2(n^*)$ is well-defined. Now, from $n^* > e$, we also get $n^*\log n^* \geq n^*$, which together with $n^* \geq 1/a$ proves that $n^*\log n^* \geq 1/a$.

To verify $q_1(n^*) \geq q_2(n^*)$ note first that from our assumptions on a and b, $2b\log\frac{1}{a} \geq 4 \geq \sqrt{e}$. Hence,

$$q_1(n^*) = 2\log(2b\log\tfrac{1}{a}) = \log(4b^2\log^2\tfrac{1}{a}) \geq \log(2b^2\log\tfrac{1}{a}) = \log b + \log(2b\log\tfrac{1}{a})$$

which holds, as by our condition on a, $\log(1/a) \geq \frac{1}{2}$. On the other hand,

$$\begin{aligned} q_2(n^*) &= \log b + \log(\log n^*) = \log b + \log\log\left(\tfrac{2}{a}\log(2b\log\tfrac{1}{a})\right) \\ &< \log b + \log\log(\tfrac{2b}{a}\log\tfrac{1}{a}) && (\log 2x < x) \\ &< \log b + \log\log(\tfrac{2b}{a^2})\,. && (\log\tfrac{1}{a} < \tfrac{1}{a}) \end{aligned}$$

Now, using again that $\log(2x) < x$,

$$\log(\tfrac{2b}{a^2}) = \log(2b) + \log\tfrac{1}{a^2} < b + \log\tfrac{1}{a^2} \leq b\log\tfrac{1}{a^2}\,,$$

where in the last inequality we also used $b \geq 2$ and $\log\frac{1}{a^2} \geq 2$ and that for $x, y \geq 2$, $x + y \leq x\frac{y}{2} + y\frac{x}{2} = xy$. Putting together all the inequalities, we obtain $q_2(n^*) < q_1(n^*)$. □

With this, we are ready to prove Theorem 1:

Proof. Let T denote the number of phases before the algorithm exits, i.e., $|A_T| > 1$ and $|A_{T+1}| = 1$. Let U denote the event that for any phase $1 \leq t \leq T$, and for any arm $i \in A_t$ that is not yet eliminated, the mean reward μ_i of arm i is within the $g(t,\delta)$ vicinity of its estimate $\hat{\mu}_i(t)$:

$$U = \{|\hat{\mu}_i(t) - \mu_i| \leq g(t,\delta) \text{ for all } (i,t) \text{ s.t. } 1 \leq t \leq T \text{ and } i \in A_t\}\,.$$

First, we will argue about the correctness and cost of the algorithm assuming that U happens and then we will show that U indeed happens with large probability.

Assume therefore that U happens. We claim that on this event, the optimal arm $i^\star = 1$ cannot be eliminated, i.e., $1 \in A_1, \ldots, A_{T+1}$. That $1 \in A_1$ holds since $A_1 = [K]$. Now, given that $1 \in A_t$ for some $1 \leq t \leq T$, we have that $\hat{\mu}_1(t) + 2g(t,\delta) \geq \mu_1 + g(t,\delta) > \max_{j\in A_t}\mu_j + g(t,\delta) \geq \max_{j\in A_t}\hat{\mu}_j(t)$, showing that $1 \in A_{t+1}$ and hence arm 1 indeed will not be eliminated.

Now, still assuming that U happens, consider bounding N. We start by asking how big t can be for a suboptimal arm $i \neq 1$ to be still included in A_{t+1}. Intuitively, if an arm is still considered as a candidate, its suboptimality "gap" Δ_i cannot be large. Indeed, defining $\hat{\mu}^*(t) = \max_{j\in A_t}\hat{\mu}_j(t)$, from $i \in A_{t+1}$ we derive

$$\begin{aligned} \Delta_i = \mu_1 - \mu_i &\leq \hat{\mu}_1(t) + g(t,\delta) - (\hat{\mu}_i(t) - g(t,\delta)) \leq \hat{\mu}^*(t) - \hat{\mu}_i(t) + 2g(t,\delta) \\ &\leq 4g(t,\delta)\,, \end{aligned}$$

where the second inequality used that $1 \in A_t$ and hence $\hat{\mu}^*(t) \geq \hat{\mu}_1(t)$, while the last inequality used that $i \in A_{t+1}$. Hence, by the definition of $\widehat{T}_i \doteq \widehat{T}_i(\delta)$, from $i \in A_{t+1}$ it follows that $t < \widehat{T}_i$. In particular, for any $t \geq \widehat{T}_i + 1$, $i \notin A_t$. As

a matter of fact, for any $i > 2$, $t \geq \widehat{T}_i + 1$, and $j \geq i$, j cannot be in A_t. Hence, $A_t \subset \{1, \ldots, i-1\}$. By reindexing, for $1 \leq i \leq K$ and using $\widehat{T}_{K+1} = 0$, we conclude that

$$t \geq \widehat{T}_{i+1} + 1 \text{ implies that } A_t \subset [i]\,, \quad 1 \leq i \leq K\,. \tag{10}$$

Since (10) implies that for $t \geq \widehat{T}_2 + 1$, A_t is a singleton, $T \geq \widehat{T}_2 + 1$ cannot hold. Hence, $T \leq \widehat{T}_2$. Now, we can bound N, the total number of probes used before termination:

$$N = \sum_{t=1}^{T} \sum_{p=1}^{P} C_p(t) = \sum_{t=1}^{T} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{IP}}(A_t, f(t))\,, \tag{11}$$

where we set $A_t = \{1\}$ for $t > T$. Now, we divide the set $\{1, \ldots, \widehat{T}_2\}$ into the disjoint intervals $S_i = \{\widehat{T}_{i+1} + 1, \ldots, \widehat{T}_i\}$, $i = 2, \ldots, K$. Using that, by (10), for any $t \in S_i$ it holds that $A_t \subset [i]$ and thus $\mathcal{C}_{\mathrm{IP}}(A_t, f(t)) \leq \mathcal{C}_{\mathrm{IP}}([i]\,, f(t))$ (where we used that for any $A \subset B$, $n \in \mathbb{N}$, $\mathcal{C}_{\mathrm{IP}}(A, n) \leq \mathcal{C}_{\mathrm{IP}}(B, n)$), we get

$$N \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} \mathcal{C}_{\mathrm{IP}}([i]\,, 2^t)\,,$$

proving (4).

It remains to lower bound the probability that U happens by $1 - \delta$. As usual, we do this by upper bounding the probability of the complementer event $U^c = \{\exists s \in [T]\,, \exists i \in A_s \text{ s.t. } |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\}$. For the sake of simplicity, let us now assume that in each phase t, for each arm in A_t, we use only the first $f(t)$ observed rewards and drop the potential "overflow". In fact, by dropping additional observations, the probability of failure can only increase, hence we may make this assumption without loss of generality.

We have

$$\begin{aligned}
\Pr(U^c) &= \Pr\left(\exists s \in [T]\,, \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\
&= \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t]\,, \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right)\,.
\end{aligned}$$

Note that $\hat{\mu}_i(s)$ is defined only when $i \in A_s$. Without loss of generality we can assume that $\hat{\mu}_i(s)$ when $i \in A_s$ is calculated based on taking the average of the first $n(s) = \sum_{q=1}^{s} f(q)$ elements of an infinite i.i.d. sequence of random variables drawn from the distribution of arm i. Hence, defining $\hat{\mu}_i(s)$ as the average of the first $n(s)$ random variables in this infinite sequence, we get a consistent extension of the definition of $\hat{\mu}_i(s)$ for arbitrary $s \geq 1$.

We have

$$\begin{aligned}
\Pr(U^c) &= \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t]\,, \exists i \in A_s, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\
&\leq \sum_{t=1}^{\infty} \Pr\left(T = t, \exists s \in [t]\,, \exists i \in [K]\,, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\
&\leq \sum_{t=1}^{\infty} \sum_{i=1}^{K} \Pr\left(T = t, \exists s \in [t]\,, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\
&\leq \sum_{i=1}^{K} \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} \Pr\left(T = t, |\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right) \\
&= \sum_{i=1}^{K} \sum_{s=1}^{\infty} \Pr\left(|\hat{\mu}_i(s) - \mu_i| > g(s, \delta)\right)\,.
\end{aligned}$$

According to Hoeffding's inequality,

$$\Pr\left(|\hat{\mu}_i(s) - \mu_i| > g(s,\delta)\right) \leq 2\exp\left(-2\sum_{t=1}^{s} f(t) g(s,\delta)^2\right) \leq 2\exp\left(-2^{s+1} \cdot \frac{\log(4Ks^2/\delta)}{2^{s+1}}\right) = \frac{\delta}{2Ks^2}$$

and hence

$$\Pr(U^c) \leq \sum_{i=1}^{K}\sum_{s=1}^{\infty} \frac{\delta}{2Ks^2} < \delta\,.$$

Thus, it remains to upper bound $\widehat{T}_i = 1 + \max\left\{t \,:\, g(t,\delta) \geq \frac{\Delta_i}{4}\right\}$:

$$\begin{aligned}
\widehat{T}_i &= 1 + \max\left\{t \,:\, g(t,\delta) \geq \frac{\Delta_i}{4}\right\} \\
&= 1 + \max\left\{t \,:\, \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}} \geq \frac{\Delta_i}{4}\right\} \\
&\leq 1 + \max\left\{\log_2 n \,:\, \sqrt{\frac{\log(4K(\log_2 n)^2/\delta)}{2n}} \geq \frac{\Delta_i}{4}\right\} \\
&\leq 1 + \log_2 \max\left\{n \,:\, \frac{\Delta_i^2}{16} n \leq \log\left(\frac{4K}{\delta \cdot \log 2}\log n\right)\right\}\,.
\end{aligned}$$

To bound the maximum above, we use Lemma 7. In our problem both $b = \frac{4K}{\delta \cdot \log 2} > 2$ and $\frac{1}{a} = \frac{16}{\Delta_i^2} > e$ satisfy the conditions in this lemma. Plugging in these values of a and b in $n^* = n^*(a,b)$, we get an upper bound of $\widehat{T}_i$ in the form of $1 + \log_2\left(\frac{32}{\Delta_i^2}\log\left(\frac{16K}{\delta\cdot\log 2}\log\frac{4}{\Delta_i}\right)\right) \leq \log_2\left(\frac{64}{\Delta_i^2}\log\left(\frac{54K}{\delta}\log\frac{4}{\Delta_i}\right)\right)$, which concludes the proof of the upper bound on $\widehat{T}_i$.

Let us now turn to proving (5). According to (11), we also have

$$N = \sum_{t=1}^{T}\sum_{p=1}^{P} C_p(t) = \sum_{t=1}^{T} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{O}}(A_t, f(t)) \leq \mathcal{G}_{LP}(O,\mathcal{P}) \sum_{t=1}^{\widehat{T}_2} \mathcal{C}_{\mathrm{LP}}(A_t, f(t))\,, \tag{12}$$

Since for $A_t \in [i]$, $\mathcal{C}_{\mathrm{LP}}(A_t, f(t)) \leq \mathcal{C}_{\mathrm{LP}}([i]\,, f(t))$ also holds,

$$\begin{aligned}
N &\leq \mathcal{G}_{LP}(O,\mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} \mathcal{C}_{\mathrm{LP}}([i]\,, 2^t) \\
&= \mathcal{G}_{LP}(O,\mathcal{P}) \sum_{i=2}^{K} \left(\sum_{t=\widehat{T}_{i+1}+1}^{\widehat{T}_i} 2^t\right) \mathcal{C}_{\mathrm{LP}}([i]\,, 1) \\
&\leq \mathcal{G}_{LP}(O,\mathcal{P}) \sum_{i=2}^{K} \left(2^{\widehat{T}_i+1} - 2^{\widehat{T}_{i+1}+1}\right) \mathcal{C}_{\mathrm{LP}}([i]\,, 1) \\
&= \mathcal{G}_{LP}(O,\mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i]\,, 1) - \sum_{i=2}^{K} 2^{\widehat{T}_{i+1}+1} \mathcal{C}_{\mathrm{LP}}([i]\,, 1)\right) \\
&\leq \mathcal{G}_{LP}(O,\mathcal{P}) \sum_{i=2}^{K} \left(2^{\widehat{T}_i+1} - 2^{\widehat{T}_{i+1}+1}\right) \mathcal{C}_{\mathrm{LP}}([i]\,, 1) \\
&\leq \mathcal{G}_{LP}(O,\mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i]\,, 1) - \sum_{i=2}^{K-1} 2^{\widehat{T}_{i+1}+1} \mathcal{C}_{\mathrm{LP}}([i]\,, 1)\right) \\
&= \mathcal{G}_{LP}(O,\mathcal{P}) \left(\sum_{i=2}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i]\,, 1) - \sum_{i=3}^{K} 2^{\widehat{T}_i+1} \mathcal{C}_{\mathrm{LP}}([i-1]\,, 1)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{G}_{LP}(O,\mathcal{P})\left(2^{\widehat{T}_2+1}\mathcal{C}_{\mathrm{LP}}([2],1)+\sum_{i=3}^{K}2^{\widehat{T}_i+1}\left(\mathcal{C}_{\mathrm{LP}}([i],1)-\mathcal{C}_{\mathrm{LP}}([i-1],1)\right)\right)\\
&\le \mathcal{G}_{LP}(O,\mathcal{P})\left(\hat{N}_2(\delta)\mathcal{C}_{\mathrm{LP}}([2],1)+\sum_{i=3}^{K}\hat{N}_i(\delta)\left(\mathcal{C}_{\mathrm{LP}}([i],1)-\mathcal{C}_{\mathrm{LP}}([i-1],1)\right)\right)\\
&= \mathcal{G}_{LP}(O,\mathcal{P})\sum_{i=2}^{K}\left(\hat{N}_i(\delta)-\hat{N}_{i+1}(\delta)\right)\mathcal{C}_{\mathrm{LP}}([i],1)\\
&= \mathcal{G}_{LP}(O,\mathcal{P})\sum_{i=2}^{K}\hat{M}_i(\delta)\mathcal{C}_{\mathrm{LP}}([i],1)
\end{aligned}
$$

where $2^{\widehat{T}_i+1} \le \frac{128}{\Delta_i^2}\log\left(\frac{54K}{\delta}\log\frac{4}{\Delta_i}\right) = \hat{N}_i(\delta)$ and $\hat{N}_{K+1}(\delta)=0$. □

A.2. Proof of Theorem 2

The proofs of our lower bounds are based on the following lemma, a specialized version of Lemma 1 of Kaufmann et al. (2015). In the lemma we need the Kullback-Leibler divergence (or relative entropy) $KL(P_1,P_2)$ of two distributions: $KL(P_1,P_2)=\int P_1(dx)\log\frac{dP_1}{dP_2}(x)$ if the Radon-Nikodym derivative $\frac{dP_1}{dP_2}$ exists and is $+\infty$ otherwise. Specializing this to two Bernoulli distributions, we get the binary relative entropy function, $d(x,y)=x\log(x/y)+(1-x)\log((1-x)/(1-y))$ defined for $x,y\in[0,1]$. (Define $d(0,0)=d(1,1)=0$, $d(0,1)=d(1,0)=+\infty$.)

Lemma 8. *Let $\hat{i}^*\in[K]$ be the arm returned by some algorithm after observing reward from arm $i\in[K]$ M_i times and let $\hat{i}^*=0$ if the algorithm never stops. For any $a\in[K]$, let U_a denote the event that $\hat{i}^*=a$. Then, for any two environments D^1 and D^2, and for any $a\in[K]$,*

$$
\sum_{i=1}^{K}\mathbb{E}_{D^1}[M_i]KL(D_i^1,D_i^2)\ge d(\Pr_{D^1}(U_a),\Pr_{D^2}(U_a))\,,
$$

where $\mathbb{E}_{D^j}$ and $\Pr_{D^j}$ denote expectation and probability, respectively, under the assumptions that the environment is D^j.

Proof. The relative entropy of two one-dimensional Gaussian distributions with common variance $\sigma^2=1/4$ and mean difference m is $m^2/(2\sigma^2)$. Let G_μ denote the Gaussian distribution with mean μ. Hence, $KL(G_\mu,G_{\mu+a})=2a^2$ for any $\mu,a\in\mathbb{R}$. Further, for any $\delta\in(0,1/2)$,

$$
d(1-\delta,\delta)=(1-\delta)\log\frac{1-\delta}{\delta}+\delta\log\frac{\delta}{1-\delta}>\frac{1}{2}\log\frac{1}{2\delta}+\delta\log\delta\ge\frac{1}{2}\log\frac{1}{2\delta}-\frac{1}{e}>\frac{1}{2}\log\frac{1}{6\delta}\,. \tag{13}
$$

Pick $\mu_1=1/2$, $\mu_i=\mu_1-\Delta_i$ and let $D^0=(D^1,\dots,D^K)\doteq(G_{\mu_1},\dots,G_{\mu_K})$. Define D^1 to be the modification of D^0 when D_2 is replaced by $G_{\mu_1+\varepsilon}$ and let D^i with $2\le i\le K$ be the modification of D^0 when D_i is replaced by $G_{\mu_1+\varepsilon}$ with some $\varepsilon>0$. As in the proof of Theorem 2 in Kaufmann et al. (2015), we apply Lemma 8 to the K pairs of environments $(D^0,D^1),\dots,(D^0,D^K)$ and arm $a=1$.

We have $KL(D_j^0,D_j^1)=0$ unless $j=2$ in which case $KL(D_2^0,D_2^1)=KL(G_{\mu_2},G_{\mu_1+\varepsilon})=(\Delta_2+\varepsilon)^2$. Also, for any $2\le i\le K$, $1\le j\le K$, $KL(D_j^0,D_j^i)=0$ unless $i=j$ in which case $KL(D_i^0,D_i^i)=KL(G_{\mu_i},G_{\mu_1+\varepsilon})=2(\Delta_i+\varepsilon)^2$. Further, the optimal arm in D^0 is arm one, while the optimal arm in D^i is arm i because $\varepsilon>0$. Hence, if U is the event that the algorithm picks arm one, then, since the algorithm is admissible, $\Pr_{D^0}(U)\ge1-\delta$, and $\Pr_{D^i}(U)\le\delta$. Combined with (13), letting M_i denote the number of observations from arm i, we get

$$
\mathbb{E}_{D^0}[M_1]\ge\frac{1}{4(\Delta_2+\varepsilon)^2}\log\frac{1}{6\delta}\,,\quad \mathbb{E}_{D^0}[M_i]\ge\frac{1}{4(\Delta_i+\varepsilon)^2}\log\frac{1}{6\delta}\,,\quad 2\le i\le K\,.
$$

Define N_p the number of times probe p is used. Then, $N=\sum_{p\in\mathcal{P}}N_p$ and $M_i=\sum_{p:i\in p}N_p$. Combining this with the previous inequalities leads to the linear program as shown in Theorem 2.

□

A.3. Proof of Theorem 3

Proof. The algorithm contains two stages: First, in the for loop, we aim to find a probe that contains an $\varepsilon/2$-optimal arm with probability at least $1-\delta/2$, in $O(|A_1|\varepsilon^{-2}\log(|\pi_{\max}|/\delta))$; then we find an arm that is $\varepsilon/2$-optimal arm within this probe with probability at least $1-\delta/2$ after $O(\varepsilon^{-2}\log(|\pi_{\max}|/\delta))$ probes.

First we will analyze the algorithm on the first stage. We need to show that

$$\Pr\left(\hat{\mu}_{\hat{\pi}^*} > \mu_{\pi^*} - \frac{\varepsilon}{2}\right) \geq 1 - \frac{\delta}{2} \tag{14}$$

where $\mu_\pi = \max_{i\in\pi}\mu_i$ for all $\pi \in A_1$ and $\pi^* = \operatorname{argmax}_\pi \mu_\pi$. Clearly, $\mu_{\pi^*} = \mu_1$, the expectation of the best arm.

Let $\pi_t = \operatorname{argmax}_{\pi\in A_t}\mu_\pi$. Let $\Pr_t$ and $\mathbb{E}_t$ denote the conditional probability and conditional expectation given all randomness before phase t. To prove (14), we will first show that

$$\Pr_t\left(\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\right) \geq 1 - \delta_t\,.$$

Define $A_t^\varepsilon = \{\pi \in A_t : \mu_\pi \leq \mu_{\pi_t} - \varepsilon_t\}$ and $A_t^* = \{\pi \in A_t : \hat{\mu}_\pi(t) > \hat{\mu}_{\pi_t}(t)\}$. Then, for any $\pi \in A_t$, the event $\{\pi \in A_t^* \cap A_t^\varepsilon\} \wedge \{\hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \varepsilon_t/2\}$ implies $\{\hat{\mu}_\pi(t) > \mu_\pi + \varepsilon_t/2\}$. Thus, for any $\pi \in A_t, \pi \neq \pi_t$,

$$\begin{aligned}
&\Pr_t\left(\pi \in A_t^* \cap A_t^\varepsilon \middle| \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \\
&\leq \Pr_t\left(\hat{\mu}_\pi(t) > \mu_\pi + \frac{\varepsilon_t}{2} \middle| A_t, \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \\
&= \Pr_t\left(\hat{\mu}_\pi(t) > \mu_\pi + \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{3},
\end{aligned}$$

where (i) the equality holds since the samples from the arms in π and π_t are independent, and (ii) the last inequality holds, since by Hoeffding's inequality (Cesa-Bianchi & Lugosi, 2006),

$$\Pr_t\left(\hat{\mu}_i(t) > \mu_\pi + \frac{\varepsilon_t}{2}\right) \leq \Pr_t\left(\hat{\mu}_i(t) > \mu_i + \frac{\varepsilon_t}{2}\right) < \frac{\delta_t}{3|\pi|}$$

for all $i \in \pi$, since $\hat{\mu}_i(t)$ is estimated from $(2/\varepsilon_t)^2\log(3|\pi|/\delta_t)$ samples, and the union bound implies that this inequality simultaneously holds for all $i \in \pi$ with probability $\delta_t/3$. Furthermore, by definition $\pi_t \notin A_t^\varepsilon$, hence $\Pr_t\left(\pi_t \in A_t^* \cap A_t^\varepsilon \middle| \hat{\mu}_{\pi_t}(t) \geq \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) = 0$. Therefore,

$$\mathbb{E}_t\left[\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|}\middle| \hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right] \leq \frac{\delta_t}{3}$$

Applying Markov's inequality, we have

$$\Pr_t\left(\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|} \geq \frac{1}{2}\middle| \hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}\right) \leq \frac{2\delta_t}{3}.$$

Note that again by Hoeffding's inequality and the union bound, $\Pr_t(\hat{\mu}_{\pi_t}(t) > \mu_{\pi_t} - \frac{\varepsilon_t}{2}) \geq 1 - \frac{\delta_t}{3}$, and $\{\frac{|A_t^* \cap A_t^\varepsilon|}{|A_t|} < \frac{1}{2}\}$ implies $\{\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\}$. Then, by union bound, we get

$$\Pr_t\left(\mu_{\pi_{t+1}} > \mu_{\pi_t} - \varepsilon_t\right) \geq 1 - \delta_t$$

Since the bound is constant, the unconditional probability also satisfies this inequality, and so, by the union bound,

$$\Pr\left(\mu_{\hat{\pi}^*} \leq \mu_{\pi^*} - \frac{\varepsilon}{2}\right) \leq \sum_{t=1}^{\log_2|A_1|} \Pr\left(\mu_{\pi_{t+1}} \leq \mu_{\pi_t} - \varepsilon_t\right) \leq \sum_{t=1}^{\log_2|A_1|} \delta_t < \frac{\delta}{2},$$

proving (14).

Next we will calculate the probe complexity until $\widehat{\pi}^*$ is found:

$$\sum_{t=1}^{\log_2|A_1|}\sum_{\pi\in A_t}\frac{4}{\varepsilon_t^2}\log\frac{3|\pi|}{\delta_t}\le\sum_{t=1}^{\log_2|A_1|}\frac{4|A_t|}{\varepsilon_t^2}\log\frac{3|\pi_{\max}|}{\delta_t}=O\left(\frac{|A_1|}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right)$$

Now we will analyze the second stage, by showing that it finds an $\varepsilon/2$-optimal arm from $\widehat{\pi}^*$ with probability at least $1-\delta/2$. Assume that the first stage ran for T phases, so we will consider conditional probabilities $\Pr_T$ conditioned on the first T phases of the first stage.

Let $i^*_{\widehat{\pi}^*}$ denote the optimal arm in $\widehat{\pi}^*$, $\hat{\mu}_i$ be the empirical mean reward for arm $i\in\widehat{\pi}^*$ in the second stage, computed from $\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$ samples, $\hat{i}^*=\operatorname{argmax}_{i\in\widehat{\pi}^*}\hat{\mu}_i$, $A^\varepsilon=\{i\in\widehat{\pi}^*:\mu_i\le\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{2}\}$ and $A^*=\{\hat{\mu}_i>\hat{\mu}_{i^*_{\widehat{\pi}^*}}\}$. Clearly, $\{\hat{\mu}_{i^*_{\widehat{\pi}^*}}\ge\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{4}\}$ and $\{\forall i\in A^\varepsilon,\hat{\mu}_i\le\mu_i+\frac{\varepsilon}{4}\}$ imply $\{|A^\varepsilon\cap A^*|=\varnothing\}$, which in turn implies $\{\hat{i}^*\notin A^\varepsilon\}$. Therefore,

$$\begin{aligned}\Pr_T\left(\mu_{\hat{i}^*}>\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{2}\right)&\ge\Pr_T\left(\hat{\mu}_{i^*_{\widehat{\pi}^*}}\ge\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{4}\wedge\forall i\in A^\varepsilon,\hat{\mu}_i\le\mu_i+\frac{\varepsilon}{4}\right)\\&\ge1-\Pr_T\left(\hat{\mu}_{i^*_{\widehat{\pi}^*}}<\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{4}\right)-\sum_{i\in A^\varepsilon}\Pr_T\left(\hat{\mu}_i>\mu_i+\frac{\varepsilon}{4}\right).\end{aligned}$$

Applying Hoeffding's inequality, we have

$$\Pr_T\left(\hat{\mu}_i-\mu_i>\frac{\varepsilon}{4}\right)\le e^{-\frac{n\varepsilon^2}{8}}=\frac{\delta}{2|\widehat{\pi}^*|}$$

where $n=\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$. Note that the same probability bound holds for $\mu_i-\hat{\mu}_i>\frac{\varepsilon}{4}$. Therefore,

$$\Pr_T\left(\mu_{\hat{i}^*}\ge\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{2}\right)\ge1-\frac{(|A^\varepsilon|+1)\delta}{2|\widehat{\pi}^*|}\ge1-\frac{\delta}{2}.$$

Since the bound is independent of the condition, we also have

$$\Pr\left(\mu_{\hat{i}^*}\ge\mu_{i^*_{\widehat{\pi}^*}}-\frac{\varepsilon}{2}\right)\ge1-\frac{\delta}{2}.$$

Combining with (14), we obtain

$$\Pr(\mu_{\hat{i}^*}\ge\mu_{i^\star}-\varepsilon)\ge1-\delta\,.$$

Finally, the total number of probes can be bounded as

$$N=O\left(\frac{|A_1|}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right)+\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}=O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right)\qquad(|A_1|=\mathcal{C}_{\mathrm{O}}([K],1))$$

□

A.4. Proof of Theorem 4

Proof. As in earlier proofs, we are going to use $\Pr_t$ and $\mathbb{E}_t$ to denote the conditional probability and conditional expectation, respectively, given all randomness before phase t, and we denote the σ-algebra corresponding to the latter by $\mathcal{F}_{t-1}$.

First we are going to bound the number of phases in running EGEWP. We start with the following simple observation: For any $i\neq1$ and t such that $T\ge t\ge\log_2\frac{1}{\Delta_i}$, $i\in A_t$ and $1\in A_t$, the event $C_{t,i}=\{\mu_{i_t}\ge\mu_1-\frac{\varepsilon_t}{2}\}\wedge\{\mu_{i_t}\le\hat{\mu}_{i_t}(t)+\frac{\varepsilon_t}{2}\}\wedge\{\mu_i\ge\hat{\mu}_i(t)-\frac{\varepsilon_t}{2}\}$ implies $i\notin A_{t+1}$. This holds since given $C_{t,i}$,

$$\hat{\mu}_{i_t}(t)\ge\mu_{i_t}-\frac{\varepsilon_t}{2}\ge\mu_1-\varepsilon_t\ge\mu_i+3\varepsilon_t\ge\hat{\mu}_i(t)+\frac{5}{2}\varepsilon_t>\hat{\mu}_i(t)+\varepsilon_t$$

where in the third step we used that $\mu_1 - \mu_i = \Delta_i \geq 2^{-t} = 4\varepsilon_t$ for $t \geq \log_2 \frac{1}{\Delta_i}$. Now assume that $\mathcal{F}_{t-1}$ is such that $1 \in A_t$ and $\pi \in \Pi_t$. Then, for any $t \geq \log_2 \frac{1}{\Delta_2}$,

$$\begin{aligned}
&\Pr_t\left(\exists i \in \pi, i \neq 1, i \in A_{t+1}\right) \\
&\leq \Pr_t\left(\mu_{i_t} < \mu_1 - \frac{\varepsilon_t}{2}\right) + \Pr_t\left(\mu_{i_t} > \hat{\mu}_{i_t}(t) + \frac{\varepsilon_t}{2}\right) + \sum_{i \in \pi, i \neq 1} \Pr_t\left(\mu_i < \hat{\mu}_i(t) - \frac{\varepsilon_t}{2}\right). \qquad (15)
\end{aligned}$$

Now, for any $t \geq 1$ and $\mathcal{F}_{t-1}$ as above,

$$\Pr_t\left(\mu_{i_t} < \mu_1 - \frac{\varepsilon_t}{2}\right) \leq \delta_t$$

by the high probability guarantee for the success of MEWP and the fact that new samples are used in each phase. Furthermore, for any $i \in \pi$ and $t \geq 1$,

$$\Pr_t\left(\mu_i < \hat{\mu}_i(t) - \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{2|\pi|} \quad \text{and} \quad \Pr_t\left(\mu_i > \hat{\mu}_i(t) + \frac{\varepsilon_t}{2}\right) \leq \frac{\delta_t}{2|\pi|} \qquad (16)$$

by Hoeffding's inequality since $\hat{\mu}_i$ is computed from $2\varepsilon_t^{-2}\log(2|\pi|/\delta_t)$ new samples. Finally, since i_t is selected based on different samples than the ones used in estimating $\hat{\mu}_{i_t}$, denoting by $\pi_t(j)$ the partition cell of A_t containing j, we have

$$\begin{aligned}
\Pr_t\left(\mu_{i_t} > \hat{\mu}_{i_t}(t) + \frac{\varepsilon_t}{2}\right) &= \sum_{j \in A_t} \Pr_t\left(\mu_j > \hat{\mu}_j(t) + \frac{\varepsilon_t}{2} | i_t = j\right) \Pr_t\left(i_t = j\right) \\
&= \sum_{j \in A_t} \Pr_t\left(\mu_j > \hat{\mu}_j(t) + \frac{\varepsilon_t}{2}\right) \Pr_t\left(i_t = j\right) \\
&\leq \sum_{j \in A_t} \frac{\delta_t}{2|\pi_t(j)|} \Pr_t\left(i_t = j\right) \leq \frac{\delta_t}{2}. \qquad (17)
\end{aligned}$$

Continuing (15) with the above inequalities, we obtain that for any $t \geq \log_2 \frac{1}{\Delta_2}$ and $\mathcal{F}_{t-1}$ such that $1 \in A_t$ and $\pi \in \Pi_t$,

$$\Pr_t\left(\exists i \in \pi, i \neq 1, i \in A_{t+1}\right) \leq \delta_t + \frac{\delta_t}{2} + \frac{\delta_t}{2} = 2\delta_t. \qquad (18)$$

Since the same $2\delta_t$ bound holds for ny $\mathcal{F}_{t-1}$ with $1 \in A_t$ and $\pi \in \Pi_t$, we also have

$$\Pr\left(\exists i \in \pi, i \neq 1, i \in A_{t+1} | \pi \in \Pi_t, 1 \in A_t\right) \leq 2\delta_t. \qquad (19)$$

Furthermore, for any $t \geq 1$, the events $\{1 \in A_t\}$, $\{\hat{\mu}_1(t) \geq \mu_1 - \frac{\varepsilon_t}{2}\}$, and $\{\hat{\mu}_{i_t}(t) \leq \mu_{i_t} + \frac{\varepsilon_t}{2}\}$ imply that $1 \in A_{t+1}$, since

$$\hat{\mu}_1(t) \geq \mu_1 - \frac{\varepsilon_t}{2} \geq \mu_{i_t} - \frac{\varepsilon_t}{2} \geq \hat{\mu}_{i_t}(t) - \varepsilon_t\,.$$

Therefore, from (16) (for $i = 1$) and (17) we get

$$\Pr(1 \in A_{t+1} | 1 \in A_t) \geq 1 - \frac{\delta_t}{2} - \frac{\delta_t}{2} = 1 - \delta_t\,. \qquad (20)$$

The above inequality shows that the optimal arm 1 is not eliminated with high probability, while (19) shows that for large enough t, the suboptimal arms are eliminated with high probability. Thus, it remains to quantify how fast the suboptimal arms are eliminated. To this end, we show that the number of probes used in every phase decays exponentially fast for $t \geq \log_2 \frac{1}{\Delta_2}$. Let $\Pi_t^+ = \{\pi \in \Pi_t : \exists i \in \pi, i \in A_{t+1}\}$. Then, for any $t \geq \log_2 \frac{1}{\Delta_2}$ and $\mathcal{F}_{t-1}$ with $1 \in A_t$, we have $\mathbb{E}_t\left[|\Pi_t^+ - \pi_t(1)|\right] \leq 2\delta_t |\Pi_t - \pi_t(1)|$ by (18), and so

$$\mathbb{E}_t\left[\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|}\right] \leq 2\delta_t\,.$$

Again, since the right hand side is independent of the conditioning in $\mathbb{E}_t$, we can replace the conditioning on $\mathcal{F}_{t-1}$ with conditioning on $1 \in A_t$; then, by Markov's inequality, for any $z > 0$ and $t \geq \log_2 \frac{1}{\Delta_2}$,

$$\Pr\left(\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|} > \frac{1}{z}\middle| 1 \in A_t\right) \leq \frac{\mathbb{E}\left[\frac{|\Pi_t^+ - \pi(1)|}{|\Pi_t - \pi_t(1)|}\right]}{z} \leq 2z\delta_t\,. \tag{21}$$

Now define the event

$$B(t) = \left\{\frac{|\Pi_t^+ - \pi_t(1)|}{|\Pi_t - \pi_t(1)|} \leq \frac{1}{z}\right\} \wedge \{\forall i \in \pi_t(1), i \neq 1, i \notin A_{t+1}\};$$

note that $\pi(1)$, and hence $B(t)$, is defined when $1 \in A_t$. Then, by (19),(21), and the union bound,

$$\Pr(B(t)|1 \in A_t) \geq 1 - 2z\delta_t - 2\delta_t = 1 - 2(z+1)\delta_t. \tag{22}$$

Next we consider when the algorithm stops if $1 \in A_t$ and $B(t)$ happen in each phase $t \geq \log_2 \frac{1}{\Delta_2}$. Note that, denoting the last phase of the algorithm by T, the probability of this event can be bounded from below as

$$\Pr\left(\{\forall t \in [T]\,, 1 \in A_t\} \wedge \{\forall \log_2 \frac{1}{\Delta_2} \leq t \leq T, B(t)\}\right) \geq 1 - \sum_{t=1}^{\infty}(2z+3)\delta_t = 1 - \sum_{t=1}^{\infty}\frac{(2z+3)\delta}{50t^3} \geq 1 - \frac{3\delta(2z+3)}{100}\,, \tag{23}$$

by (20), (22), the union bound, and since $\sum_{t=1}^{\infty} 1/t^3 < 1 + \int_1^{\infty} 1/t^3 dt = 3/2$.

If $z > 1$ and $|\Pi_t| \leq z$, then

$$|\Pi_t^+ - \pi(1)| \leq \frac{|\Pi_t - \pi(1)|}{z} \leq \frac{z-1}{z} < 1\,,$$

which means that $\Pi_t^+ \subset \{\pi(1)\}$. Also, $B(t)$ implies that all suboptimal arms in $\pi(1)$ are eliminated, which leads to the fact that only the optimal arm 1 can survive after phase t. According to the algorithm, there must be at least one arm left after the elimination of each phase, so we can conclude that if for some $z > 1$ and phase $t > \log_2 \frac{1}{\Delta_2}$, $\{1 \in A_t\}$, $|\Pi_t| \leq z$, and $B(t)$ holds, the algorithm must stop after this phase and return the optimal arm $i^\star = 1$.

If $|\Pi_t| > z$, and $\{1 \in A_t\}$ and $B(t)$ holds, then

$$\frac{|\Pi_t^+|}{|\Pi_t|} \leq \frac{|\Pi_t^+ - \pi(1)| + 1}{|\Pi_t|} \leq \frac{|\Pi_t - \pi(1)| + z}{z|\Pi_t|} = \frac{|\Pi_t| + z - 1}{z|\Pi_t|} \leq \frac{(z-1)+(z+1)}{z(z+1)} = \frac{2}{z+1}\,,$$

Since repartitioning in the next phase will not increase the number of probes needed to cover A_{t+1} compared to Π_t^+, we have $|\Pi_{t+1}| \leq |\Pi_t^+|$. Therefore, for any $z > 1$ and $t \geq \log_2 \frac{1}{\Delta_2}$ such that $\{1 \in A_t\} \wedge B(t)$ holds, $|\Pi_t| > z$, implies

$$\frac{|\Pi_{t+1}|}{|\Pi_t|} \leq \frac{2}{z+1}\,. \tag{24}$$

For simplicity, we choose $z = 15$. Then, by (23), the probability of the event $\{\forall t \in [T]\,, 1 \in A_t\} \wedge \{\forall \log_2 \frac{1}{\Delta_2} \leq t \leq T, B(t)\}$ is at least $1-\delta$; thus, it is enough to bound the probe complexity of the algorithm under the latter event. Assuming the event holds, by the choice of z we have that after $t \geq \log_2 \frac{1}{\Delta_2}$ phases, $|\Pi_t| \geq 16$ implies $\frac{|\Pi_{t+1}|}{|\Pi_t|} \leq \frac{1}{8}$, and the algorithm stops after phase t if $|\Pi_t| \leq 15$. Let $s = \log_2 \frac{1}{\Delta_2}$. Then the algorithm must run into one of the following three cases: (a) $T < s$, (b) $T \geq s$ and $|\Pi_t| \geq 16$ for $s \leq t \leq T$, (c) $T \geq s$ and $|\Pi_t| \geq 16$ for $s \leq t \leq T-1$, $|\Pi_T| \leq 15$.

Here we only consider the last two cases where $T \geq s$; the upper bound obtained this way trivially hold for case (a), as well. For $T \geq s$, we divide the T phases into two parts: $1 \leq t < s$ and $s \leq t \leq T$. In the second part, by definition, $|\Pi_t| \geq 16$ for $s \leq t \leq T-1$, and so $|\Pi_t| \leq \mathcal{C}_{\mathrm{O}}([K]\,, 1)\left(\frac{1}{8}\right)^{t-s}$ for $s \leq t \leq T$ by (24). Therefore, the probe complexity of the algorithm, without the samples used by the MEWP subroutine, is

$$\sum_{t=1}^{s-1}\sum_{\pi\in\Pi_t}\frac{2}{\varepsilon_t^2}\log\frac{2|\pi|}{\delta_t} + \sum_{t=s}^{T}\sum_{\pi\in\Pi_t}\frac{2}{\varepsilon_t^2}\log\frac{2|\pi|}{\delta_t}$$

$$\leq 32\mathcal{C}_{\mathrm{O}}([K],1)\sum_{t=1}^{s-1}4^t\log\frac{100|p_{\max}|t^3}{\delta}+32\mathcal{C}_{\mathrm{O}}([K],1)\sum_{r=0}^{T-s}\left(\frac{1}{8}\right)^r 4^{r+s}\log\frac{100|p_{\max}|(r+s)^3}{\delta}\,.$$

Here the first term on the right hand side is clearly bounded from above by

$$C_1\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)$$

for some universal positive constant C_1, while the second term can be bounded as

$$C_2\cdot\mathcal{C}_{\mathrm{O}}([K],1)4^s\left(\sum_{r=0}^{T-s}\frac{1}{2^r}\log\frac{s\cdot|p_{\max}|}{\delta}+\sum_{r=0}^{T-s}\frac{\log r}{2^r}\right)\leq C_3\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right)\qquad(r+s\leq rs)$$

for universal contants $C_2, C_3 > 0$. In conclusion, the total probe complexity without the samples used by median elimination is

$$O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right)\,.$$

The last thing is to show that the probe complexity of the MEWP subroutine is dominated by the above quantity. To show this, consider each phase t, the number of probes used outside median elimination is $\sum_{\pi\in\Pi_t}\frac{2}{\varepsilon_t^2}\log\frac{2|\pi|}{\delta_t}$ which is relaxed to $\frac{2\mathcal{C}_{\mathrm{O}}(A_t,1)}{\varepsilon_t^2}\log\frac{2|p_{\max}|}{\delta_t}$ in our analysis. According to Theorem 3, MEWP in phase t uses $O\left(\frac{\mathcal{C}_{\mathrm{O}}(A_t,1)}{\varepsilon_t^2}\log\frac{|\pi_{\max}|}{\delta_t}\right)$ probes, where $|\pi_{\max}|=\max_{\pi\in\Pi_t}|\pi|\leq|p_{\max}|$. So taking the probe complexity of median elimination processes into account we still have the total probe complexity as

$$N=O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right)\,.$$

□

A.5. Proof of Theorem 5

Proof. Let T denote the number of phases that the algorithm runs until the stopping condition is satisfied and U denote the event that all confidence bounds hold throughout the process:

$$U=\{|\hat{\mu}_i(t)-\mu_i|\leq g(t,\delta)\text{ for all }(i,t)\text{ s.t. }1\leq t\leq T\text{ and }i\in A_t\}\,.$$

In the proof of Theorem 1, we have already shown that $\Pr(U)\geq 1-\delta$. So the remaining of the proof contains two parts given the fact that U holds: (i) if T is finite thus $\hat{S}^*$ is returned, each arm in $\hat{S}^*$ must be (ε,m)-optimal, and (ii) the probe complexity is upper bounded.

First we will show that if $T<\infty$ then each arm $i\in\hat{S}^*$ must be (ε,m)-optimal. Since $\hat{S}^*=A_T^a\cup H_T$, if $i\in\hat{S}$, i belongs to either A_T^a or H_T. If $i\in A_T^a$, we use the following proposition to show that $1\leq i\leq m$.

Proposition 9. *If U holds, then for any $2\leq t\leq T$, if $i\in A_t^a$ then $1\leq i\leq m$, if $i\in A_t^r$ then $m+1\leq i\leq K$.*

Proof. For $t=2$, if $i\in A_2^a$, then $\hat{\mu}_i(1)>\max_{j\in L_1}\hat{\mu}_j(1)+2g(1,\delta)$. Since $|L_1|=K-m$, we can find at least $K-m$ arms such that for each j of them $\hat{\mu}_i(1)>\hat{\mu}_j(1)+2g(1,\delta)$. From U we know that $\mu_i>\mu_j$, which means there are at least $K-m$ arms worse than i, hence $1\leq i\leq m$ holds. On the other hand, if $i\in A_T^r$, for similar reason, we can find at least m arms better than i and thus $m+1\leq i\leq K$. Next we will show that if it holds for t then it also holds for $t+1$.

If it holds for t, then we have $\#\{i:1\leq i\leq m, i\in A_t\}=m-|A_t^a|$ and $\#\{i:m+1\leq i\leq K, i\in A_t\}=K-m-|A_t^r|$. For $i\in A_t$ and $i\in A_{t+1}^a$, we can find at least $|L_t|=K-m-|A_t^r|$ arms in A_t worse than i so $1\leq i\leq m$ must hold. Similarly, for $i\in A_t$ and $i\in A_{t+1}^r$, we can find at least $|H_t|=m-|A_t^a|$ arms in A_t better than i so $m+1\leq i\leq K$ must hold. Then by induction, Proposition 9 holds. □

We now continue the proof of Theorem 5. Proposition 9 shows that if $i \in A_T^a$ then $1 \leq i \leq m$. For the other case, if $i \in H_T$, then $\hat{\mu}_i(T) \geq \max_{j \in L_T} \hat{\mu}_j(T) + 2g(T,\delta) - \varepsilon$. Next we will show that $\mu_i \geq \mu_m - \varepsilon$ by discussing the following two cases:

If $1 \leq i \leq m$ then $\mu_i \geq \mu_m - \varepsilon$ must hold. If $m+1 \leq i \leq K$, since $i \in H_T$ and all arms in A_T^r must be $K-m$ worst, then there exists $1 \leq j \leq m$ such that $j \in L_T$ and thus $\hat{\mu}_i(T) \geq \hat{\mu}_j(T) + 2g(T,\delta) - \varepsilon$. Therefore $\mu_i \geq \mu_j - \varepsilon \geq \mu_m - \varepsilon$.

Now we have shown that if $T < \infty$, every arm in $\hat{S}^* = A_T^a \cup H_T$ must be (ε, m)-optimal. Next we will prove that if U holds then the probe complexity is upper bounded by the following propositions.

Proposition 10. *For $1 \leq t < T$, $g(t,\delta) > \varepsilon/2$.*

Proof. If $g(t,\delta) \leq \varepsilon/2$, then $\min_{i \in H_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) + 2g(t,\delta) - \varepsilon$. The stopping condition is satisfied, thus $T = t$. □

Proposition 11. *For $1 \leq t < T$, if $i \in A_{t+1}$, then $g(t,\delta) \geq (\mu_i - \mu_{m+1})/4$ if $1 \leq i \leq m$, and $g(t,\delta) \geq (\mu_m - \mu_i)/4$ if $m+1 \leq i \leq K$.*

Proof. For $i \in A_t$, $1 \leq i \leq m$, if $g(t,\delta) < (\mu_i - \mu_{m+1})/4$, since $\#\{i : m+1 \leq i \leq K, i \in A_t\} = K - m - |A_t^r|$, there exist at least $K - m - |A_t^r|$ arms in A_t such that for each j of them $\mu_i - \mu_j > 4g(t,\delta)$. Then

$$\hat{\mu}_i(t) - \hat{\mu}_j(t) \geq (\mu_i - g(t,\delta)) - (\mu_j + g(t,\delta)) = \mu_i - \mu_j - 2g(t,\delta) > 2g(t,\delta)\,.$$

Given the fact that L_t contains $K - m - |A_t^r|$ arms with the lowest $\hat{\mu}_j(t)$s for $j \in A_t$, we have $\hat{\mu}_i(t) > \max_{j \in L_t} \hat{\mu}_j(t) + 2g(t,\delta)$, which means i must be accepted to A_{t+1}^a thus $i \notin A_{t+1}$.

Similarly, we can prove that for $i \in A_t$, $m+1 \leq i \leq K$, if $g(t,\delta) < (\mu_m - \mu_i)/4$, then i must be rejected to A_{t+1}^r. Now Proposition 11 has been proved. □

Combining Propositions 10 and 11 and the definition of $\Delta_i^{(\varepsilon,m)}$ we get that for $1 \leq t < T$, if $i \in A_{t+1}$, $g(t,\delta) \geq \Delta_i^{(\varepsilon,m)}/4$. Then following the proof of Theorem 1 gives the result of Theorem 5.

□

A.6. Proof of Theorem 6

Proof. Let T denote the number of phases that the algorithm runs until the stopping condition is satisfied and U denote the event that all confidence bounds hold throughout the process:

$$U = \{|\hat{\mu}_i(t) - \mu_i| \leq g(t,\delta) \text{ for all } (i,t) \text{ s.t. } 1 \leq t \leq T \text{ and } i \in A_t\}\,.$$

We have $\Pr(U) \geq 1 - \delta$. Similar with the proof of Theorem 5, the remaining of the proof contains two parts given the fact that U holds: (i) if T is finite thus $\hat{S}^*$ is returned, the aggregate regret of $\hat{S}^*$ must be ε-optimal, and (ii) the probe complexity is upper bounded.

First we will show that if $T < \infty$ then $\frac{1}{m}\left(\sum_{i \in [m]} \mu_i - \sum_{i \in \hat{S}^*} \mu_i\right) \leq \varepsilon$. Recall that $\hat{S}^* = A_T^a \cup H_T$. The arms in A_T^a incur no regret since Proposition 9 still holds and says that $A_T^a \subset [m]$. So we only need to show that

$$\sum_{i \in [m] \setminus A_T^a} \mu_i - \sum_{i \in H_T} \mu_i \leq m\varepsilon\,.$$

Furthermore, it is equivalent to show

$$\sum_{i \in [m] \setminus A_T^a \setminus H_T} \mu_i - \sum_{i \in H_T \setminus [m]} \mu_i \leq m\varepsilon\,.$$

Recall the stopping condition

$$\sum_{i \in H_T \setminus H_T'} (\hat{\mu}_i(T) - g(T,\delta)) \geq \sum_{i \in H_T' \setminus H_T} (\hat{\mu}_i(T) + g(T,\delta)) - m\varepsilon\,.$$

To show that the stopping condition is sufficient, we introduce some new notations:

Consider the sequence of arms in A_t sorted by their $\hat{\mu}_i(t)$, let $a_t(i)$ be the arm at the i-th position. Let

$$b_t = \max\left\{a \in \mathbb{N} : \hat{\mu}_{a_t(m_t-a+1)} - \hat{\mu}_{a_t(m_t+a)} < 2g(t,\delta)\right\},$$

where $m_t = m - |A_t^a|$.

According to the construction of H_t' we know that $H_t = \{a_t(1), ..., a_t(m_t)\}$, $H_t' = \{a_t(1), ..., a_t(m_t - b_t), a_t(m_t + 1), ..., a_t(m_t + b_t)\}$ and $|H_t \setminus H_t'| = |H_t' \setminus H_t| = b_t$.

Next we construct a set of pairs $Pair_T = \{(i,j)\}$ for $i \in [m] \setminus A_T^a \setminus H_T$ and $j \in H_T \setminus [m]$ as follows: sort $H_T \setminus [m]$ and $[m] \setminus A_T^a \setminus H_T$ both in descending order according to their $\hat{\mu}_i(T)$s (this is valid since $[m] \setminus A_T^a \subset A_T$ by Proposition 9) , then take last of $i \in [m] \setminus A_T^a \setminus H_T$ and the first $j \in H_T \setminus [m]$ as a pair into $Pair_T$, then repeat this procedure until no arm remains (Note that $|\,[m] \setminus A_T^a \setminus H_T| = |H_T \setminus [m]\,|$). Since for each pair (i,j), $i \notin H_T$ and $j \in H_T$, we have $\hat{\mu}_j(T) \geq \hat{\mu}_i(T)$.

Then

$$\begin{aligned}
&\sum_{i \in [m] \setminus A_T^a \setminus H_T} \mu_i - \sum_{i \in H_T \setminus [m]} \mu_i \\
&\leq \sum_{(i,j) \in Pair_T} (\hat{\mu}_i(T) - \hat{\mu}_j(T) + 2g(T,\delta)) \\
&\leq \sum_{(i,j) \in Pair_T^+} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)))
\end{aligned}$$

where $Pair_T^+ = \{(i,j) \in Pair_T : \hat{\mu}_j(T) - \hat{\mu}_i(T) < 2g(T,\delta)\}$. Then we will show $|Pair_T^+| \leq b_T$. This is because, if $|Pair_T^+| > b_T$, then there must be a pair $(i,j) \in Pair_T^+$ such that $j \in H_T \cap H_T'$ and $i \notin H_T \cup H_T'$. Thus $\hat{\mu}_{a_T(m_T - b_T)}(T) - \hat{\mu}_{a_T(m_T + b_T + 1)}(T) \leq \hat{\mu}_j(T) - \hat{\mu}_i(T) < 2g(T,\delta)$ which contradicts the definition of b_t.

Next we construct another set of $|Pair_T^+|$ pairs (i,j) between $i \in H_T' \setminus H_T$ and $j \in H_T \setminus H_T'$ in the similar fashion: Let

$$Pair_T' = \left\{(a_T(m_T + |Pair_T^+|), a_T(m_T - |Pair_T^+| + 1)), ..., (a_T(m_T + 1), a_T(m_T))\right\}.$$

If we consider the pairs in $Pair_T^+$ and $Pair_T'$ in the order that they are constructed, then for each corresponding $(i,j) \in Pair_T^+$ and $(i',j') \in Pair_T'$, we have $\hat{\mu}_j(T) - \hat{\mu}_i(T) \geq \hat{\mu}_{j'}(T) - \hat{\mu}_{i'}(T)$. Therefore,

$$\begin{aligned}
\sum_{i \in [m] \setminus A_T^a \setminus H_T} \mu_i - \sum_{i \in H_T \setminus [m]} \mu_i &\leq \sum_{(i,j) \in Pair_T^+} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T))) \\
&\leq \sum_{(i,j) \in Pair_T'} (2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)))
\end{aligned}$$

Consider the remaining pairs (i,j) between $i \in H_T' \setminus H_T$ and $j \in H_T \setminus H_T'$ which are not in $Pair_T'$, $2g(T,\delta) - (\hat{\mu}_j(T) - \hat{\mu}_i(T)) > 0$ still holds. Then we have

$$\sum_{i \in [m] \setminus A_T^a \setminus H_T} \mu_i - \sum_{i \in H_T \setminus [m]} \mu_i \leq \sum_{i \in H_T' \setminus H_T} (\hat{\mu}_i(T) + g(T,\delta)) - \sum_{i \in H_T \setminus H_T'} (\hat{\mu}_i(T) - g(T,\delta)) \leq m\varepsilon\,.$$

Now we have proved that the aggregate regret of $\hat{S}^*$ is ε-optimal. The remaining task is to upper bound the probe complexity.

Proposition 12. *For $1 \leq t < T$, $g(t,\delta) > m\varepsilon/4b$, where*

$$b = \max\left\{a \in \mathbb{N}^+ : \mu_{m-a+1} - \mu_{m+a} \leq \frac{m\varepsilon}{a}\right\},$$

or $b = 1$ if $\mu_m - \mu_{m+1} > m\varepsilon$.

Proof. The proposition is proved by showing that if $g(t,\delta) \leq m\varepsilon/4b$, the stopping condition must be satisfied after this phase. Recall the definition of $b_t = |H_t \setminus H'_t| = |H'_t \setminus H_t|$, we will first show that $b_t \leq b$.

If $b = \min\{m, K-m\}$, $b_t \leq b$ must hold. Next we discuss the case when $b < \min\{m, K-m\}$: Since $\mu_{m-b} - \mu_{m+b+1} > m\varepsilon/b \geq 4g(t,\delta)$, for any $1 \leq i \leq m-b$ and $m+b+1 \leq j \leq K$, if $i, j \in A_t$, then $\hat{\mu}_i(t) - \hat{\mu}_j(t) \geq \mu_i - \mu_j - 2g(t,\delta) > 2g(t,\delta)$. So there are at least $m - |A_t^a| - b = m_t - b$ arms in A_t such that for each i of them $1 \leq i \leq m-b$, as well as at least $|A_t| - m_t - b$ arms such that for each j of them $m+b+1 \leq j \leq K$. Since for each pair of such i, j, $\hat{\mu}_i(t) > \hat{\mu}_j(t)$, if $b_t > b$ then there must exist $1 \leq i \leq m-b$ and $m+b+1 \leq j \leq K$ such that $i, j \in (H_t \setminus H'_t) \cup (H'_t \setminus H_t)$. This is impossible because

$$\hat{\mu}_{a_t(m_t - b_t + 1)}(t) \geq \hat{\mu}_i(t) > \hat{\mu}_j(t) + 2g(t,\delta) \geq \hat{\mu}_{a_t(m_t+b_t)}(t) + 2g(t,\delta)\,,$$

which contradicts the definition of b_t. Hence $b_t \leq b$ holds.

Then

$$\begin{aligned}
&\sum_{i \in H'_t \setminus H_t} (\hat{\mu}_i(t) + g(t,\delta)) - \sum_{i \in H_t \setminus H'_t} (\hat{\mu}_i(t) - g(t,\delta)) \\
&= 2b_t g(t,\delta) + \sum_{i \in H'_t \setminus H_t} \hat{\mu}_i(t) - \sum_{i \in H_t \setminus H'_t} \hat{\mu}_i(t) \\
&\leq 2b_t g(t,\delta) \leq 2b_t \leq 2b \cdot \frac{m\varepsilon}{4b} \\
&\leq m\varepsilon\,,
\end{aligned}$$

which shows that the stopping condition is satisfied and thus the proposition holds. □

Note that Proposition 11 still holds here, together with Proposition 12 we get that for $1 \leq t < T$, if $i \in A_{t+1}$, $g(t,\delta) \geq \Delta_i^{(m\varepsilon/b, m)}/4$. Then following the proof of Theorem 1 gives the result of Theorem 6.

□

B. Discussion for the MEWP Algorithm

The median elimination algorithm (ME) of Even-Dar et al. (2002) works as follows: The algorithm runs in phases. In every phase t each potentially good arm is sampled $4\varepsilon_t^{-2}\log(3/\delta_t)$ times, where $\delta_t = \delta 2^{-t-1}$ and $\varepsilon_t = \varepsilon(3/4)^t/3$; then the lower half of the arms with inferior performance is eliminated, and the next phase is run with the remaining arms only. The algorithm terminates when a single arm remains.

B.1. Compared with a Naive Modification of ME

A tempting approach to address our problem would be, instead of sampling each remaining arm n times in one phase, we sample a set of probes that is a minimum n-cover of those arms. We will call this naive modification of the median elimination algorithm the naive-ME algorithm. While "naive-ME" preserves the same $O(K\varepsilon^{-2}\log(1/\delta))$ performance in the bandit case, the following proposition shows that in the full information case this algorithm requires $K^{1/2}$-times more probes than expected.

Proposition 13. *In the full information case where $\mathcal{P} = \{[K]\}$, the probe complexity of the naive-ME algorithm is at least*

$$\Omega\left(\frac{K^{1/2}}{\varepsilon^2}\log\frac{K}{\delta}\right).$$

Intuitively, the presence of the $K^{1/2}$ term is not expected since the full information case gives K times more information than the bandit case.

B.2. Further Analysis of MEWP

It can be shown that the worst case upper bound in Theorem 3 is unimprovable in both the bandit and full information setting by the following theorem.

Theorem 14. *In the full information case, for every $K \geq 2$, $\varepsilon > 0$ and $\delta \in (0, 1/2)$, and for any algorithm that returns an ε-optimal arm with probability at least $1-\delta$, there exist reward distributions $(D_1, \ldots, D_K)$ such that*

$$\mathbb{E}[N] \geq \frac{1}{16\varepsilon^2}\log\frac{K}{12\delta} \tag{25}$$

where N is the total number of probes used by the algorithm.

Moreover, for any general observation structure $\mathcal{P}$, a lower bound is

$$\mathbb{E}[N] \geq \frac{\mathcal{C}_{\mathrm{LP}}([K], 1)}{16\varepsilon^2}\log\frac{1}{6\delta}. \tag{26}$$

Compared to the upper bound of Theorem 3 in general cases, lower bound (26) has a $|\pi_{\max}|$ gap inside the $\log$ term. However, (26) is not tight since in the full information case we have a tighter lower bound $\Omega(\varepsilon^{-2}\log(K/\delta))$ in (25). Therefore, although whether the $|\pi_{\max}|$ term is tight or not is still an open question there has to be some quantity between 1 and K in the $\log$ term. Note that MEWP may not be the best choice for only finding an ε-optimal arm in practice since it does not provide distribution dependent performance. However, the worst case upper bound is theoretically good enough (has a better $\log$ term) for being a subroutine of our later algorithm EGEWP.

B.3. Empirical Comparison

We compare our MEWP algorithm to naive-ME, the naive modification of the median elimination algorithm described above. We consider the case where every subset of size $|p|$ is a probe in $\mathcal{P}$ and compare the two algorithms with different $|p|$ values when $K = 1000$. The other parameters are set as follows: $\delta = 0.1$, $\varepsilon = 0.2$, the reward distributions are all Gaussian with variance $1/4$ and means $\mu_1 = 1$, $\mu_i = 1 - (i/K)^{0.5}$ for $i \neq 1$.

The results, presented in Figure 2, show that the probe complexity of MEWP decreases faster than that of naive-ME as $|p|$ grows to K, and finally becomes much better close to the full information case ($|p| = K$). This is consistent with our theoretical results in Proposition 13 and Theorem 3 since in the bandit case both algorithms require $O(K\varepsilon^{-2}\log(1/\delta))$ probes, while in the full information case our lower bound for the naive-ME in Proposition 13 shows a $\sqrt{K}$ disadvantage compared to the upper bound of MEWP in Theorem 3.

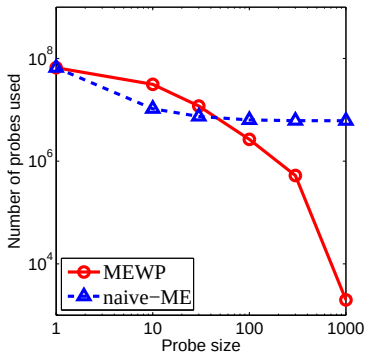

Figure 2. Comparison between MEWP and naive-ME

B.4. Proof of Proposition 13

The median elimination algorithm deterministically runs $\lceil \log_2 K \rceil$ phases since it eliminates half of the arms in each phase. In phase t, the algorithm collects $\frac{4}{\varepsilon_t^2} \log \frac{3}{\delta_t}$ samples for each arm in the set of arms A_t considered, where $\varepsilon_t = \frac{\varepsilon}{3}\left(\frac{3}{4}\right)^t$ and $\delta_t = \frac{\delta}{2^{t+1}}$, and then selects A_{t+1} to contain half of the arms with better estimated mean rewards. Under the full information setting, there is only one probe that covers all arms, so the algorithm uses that probe the probe $\frac{4}{\varepsilon_t^2} \log \frac{3}{\delta_t}$ times in each phase. Then the total probe complexity N is

$$
\begin{aligned}
N &= \sum_{t=1}^{\lceil \log_2 K \rceil} \frac{4}{\varepsilon_t^2} \log \frac{3}{\delta_t} = \sum_{t=1}^{\lceil \log_2 K \rceil} \frac{36}{\varepsilon^2} \left(\frac{16}{9}\right)^t \log \frac{6 \cdot 2^t}{\delta} \\
&\geq \frac{36}{\varepsilon^2} \left(\frac{16}{9}\right)^{\log_2 K} \log \frac{6K}{\delta} && \text{(only take the last term)} \\
&= \frac{36}{\varepsilon^2} K^{\log_2 \frac{16}{9}} \log \frac{6K}{\delta} > \frac{36}{\varepsilon^2} K^{1/2} \log \frac{6K}{\delta} \\
&= \Omega\left(\frac{K^{1/2}}{\varepsilon^2} \log \frac{K}{\delta}\right) .
\end{aligned}
$$

B.5. Proof of Theorem 14

We prove (25) and (26) separately.

First we prove a modification of Lemma 8 for the full information case.

Lemma 15. *Consider the full information case. Let $\hat{i}^* \in [K]$ be the arm returned by some algorithm after N trials if the algorithm stops and let $\hat{i}^* = 0$ if the algorithm never stops. Furthermore, for any $a \in [K]$, let U_a denote the event that $\hat{i}^* = a$. Then, for any two environments D^1 and D^2, and for any $a \in [K]$,*

$$
\mathbb{E}_{D^1}[N] \sum_{i=1}^{K} KL(D_i^1, D_i^2) \geq d(\Pr_{D^1}(U_a), \Pr_{D^2}(U_a)),
$$

where $\mathbb{E}_{D^j}$ and $\Pr_{D^j}$ denote expectation and probability under the assumptions that the environment is D^j, $KL(D_i^1, D_i^2)$ denotes the relative entropy (or Kullback-Leibler divergence) between D_i^1 and D_i^2 for all $i \in [K]$, and $d(x, y) = x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is the binary relative entropy.

Proof. Assume that the full information algorithm is applied in the bandit case in a naive way: trying each arm once in the bandit case when it would choose to try the only probe in the full information case. Then $N_1 = \ldots = N_K = N$, and the statement of the lemma follows immediately from Lemma 8. □

Proof of (25). We prove the theorem by applying Lemma 15. In order to do so, we need to construct the environments D^1 and D^2. We assume that for any $i \in [K], k \in \{1,2\}$, D_i^k is Gaussian with mean μ_i^k and variance $\sigma^2 = 1/4$. In D_2 we set $\mu_1^2 = \varepsilon, \mu_2^2 = ... = \mu_K^2 = 0$. Now consider any algorithm that returns an ε-optimal arm with probability $1-\delta$. Then we have $\Pr_{D^2}(U_1) \geq 1-\delta$. Furthermore, $\sum_{i=2}^K \Pr_{D^2}(U_i) < \delta$, and so there exists some $j \in \{2,\dots,K\}$ such that $\Pr_{D^2}(U_j) < \delta/(K-1)$. We use this j to select the expected values of the distributions D_i^1: in particular, we let $\mu_1 = \varepsilon, \mu_j = 2\varepsilon$, and $\mu_i = 0$ for all other i. Then we have $\Pr_{D^1}(U_j) \geq 1-\delta$.

Since the relative entropy of two 1-dimensional Gaussian distributions with common variance σ^2 and mean difference m is $m^2/(2\sigma^2)$, we have $\sum_{i=1}^K KL(D_i^1, D_i^2) = KL(D_j^1, D_j^2) = (2\varepsilon)^2/(2\sigma^2) = 8\varepsilon^2$. Furthermore, by the monotonicity properties of the binary entropy function d, and since $\Pr_{D^2}(U_j) < \delta/(K-1) < 1-\delta \leq \Pr_{D^1}(U_j)$, we have $d(\Pr_{D^1}(U_j), \Pr_{D^2}(U_j)) \geq d(1-\delta, \delta/(K-1))$. Thus, applying Lemma 15, we get

$$\mathbb{E}_{D^1}[N] \geq \frac{d\left(1-\delta, \frac{\delta}{K-1}\right)}{8\varepsilon^2} \,. \tag{27}$$

The last step is to bound $d\left(1-\delta, \frac{\delta}{K-1}\right)$ from below:

$$\begin{aligned} d\left(1-\delta, \frac{\delta}{K-1}\right) &= (1-\delta)\log\frac{1-\delta}{\frac{\delta}{K-1}} + \delta\log\frac{\delta}{1-\frac{\delta}{K-1}} \\ &> \frac{1}{2}\log\frac{K-1}{2\delta} + \delta\log\delta \geq \frac{1}{2}\log\frac{K-1}{2\delta} - \frac{1}{e} \\ &> \frac{1}{2}\log\frac{K-1}{6\delta} \geq \frac{1}{2}\log\frac{K}{12\delta} \,. \end{aligned}$$

Combined with (27) we have $\mathbb{E}_{D^1}[N] \geq \frac{1}{16\varepsilon^2}\log\frac{K}{12\delta}$, which concludes the proof. □

Proof of (26). Let D be an environment such that $D_i, i \in [K]$ is Gaussian with mean μ_i and variance $\sigma^2 = 1/4$, where $\mu_1 = \varepsilon$ and $\mu_i = 0$ for all $i \neq 1$.

We create K environments, $D^1, \dots, D^K$, such that D_i^k is Gaussian with mean μ_i^k and variance $\sigma^2 = 1/4$, and use Lemma 8 to lower bound the number of trials needed in environment D. For D^1, let $\mu_1^1 = -\varepsilon$ and $\mu_i^1 = \mu_i$ for all $i \neq 1$. For D^k, $k \neq 1$, let $\mu_k^k = 2\varepsilon$ and $\mu_j^k = \mu_j$ for all $j \neq k$.

Consider an algorithm A that, with probability at least $1-\delta$, returns an ε-optimal solution (in any environment satisfying the assumptions of this paper). Then, using the notation of Lemma 8, we have $\Pr_{D^k}(U_1) < \delta$ for all $k \in [K]$ and $\Pr_D(U_1) \geq 1-\delta$.

Let N_i be the number of samples observed by algorithm A for arm i. Similarly to the proof of Lemma 15, we construct a bandit algorithm from A using probes in such a way that whenever A decides to try a probe p in the original problem, the bandit version tries each arm $i \in p$ once in the bandit problem. Then the number of samples for each arm i will be the same in the original and in the bandit problem, and so, similarly to the proof of Theorem 14, Lemma 8 implies that

$$\mathbb{E}_D[N_i] \geq \frac{d(1-\delta, \delta)}{8\varepsilon^2}$$

for all $i \in [K]$. Using the derivation in Theorem 2, we get $d(1-\delta,\delta) > \frac{1}{2}\log\frac{1}{6\delta}$. Therefore, we have

$$\mathbb{E}_D[N_i] = \sum_{p \ni i} \mathbb{E}_D[N_p] \geq \frac{1}{16\varepsilon^2}\log\frac{1}{6\delta}$$

where N_p is the number of times that probe p is played. Since $\mathbb{E}_D[N] = \sum_{p\in\mathcal{P}} \mathbb{E}_D[N_p]$, lower bounding $\mathbb{E}_D[N]$ leads to

$$\mathbb{E}_D[N] \geq \mathcal{C}_{\mathrm{LP}}\left([K], \frac{1}{16\varepsilon^2}\log\frac{1}{6\delta}\right) = \frac{\mathcal{C}_{\mathrm{LP}}([K],1)}{16\varepsilon^2}\log\frac{1}{6\delta} \,. \tag{28}$$

□

C. Experimental Results

C.1. Comparing SEWP and EGEWP

To compare the performance between SEWP and EGEWP algorithms, we investigate the performance under three different probe settings: (a) the bandit case; (b) the full information case; and (c) an intermediate case where every subset of size $|p| = \sqrt{K}$ is a probe. For each scenario we consider two environments: (a) an *easy* problem where $\mu_1 = 0.3$ and $\mu_2 = ... = \mu_K = 0$ and (b) a *hard* problem where $\mu_1 = 1$ and $\mu_i = 1 - (i/n)^{0.5}$ for $i \neq 1$. Each reward distribution is Gaussian with variance $\sigma^2 = 1/4$. Under each combination of probe and distribution settings, we test the sample complexity for different values of K with $\delta = 0.1$. In the experiments we report average probe usage over 100 runs. The results are shown in Figure 6.

The results show that EGEWP performs worse than the SEWP in all settings considered, despite its favorable asymptotic performance guarantees. This phenomenon is supported by the experimental studies by Jamieson et al. (2014) in the bandit case, in which the exponential gap elimination algorithm of Karnin et al. (2013) is shown to be worse than the successive elimination algorithm of Even-Dar et al. (2002).

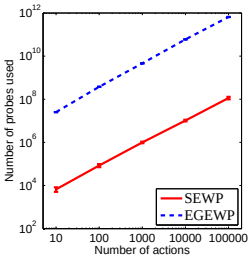

Figure 3. $|p| = 1$, easy case

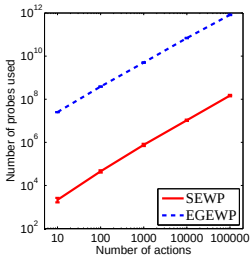

Figure 4. $|p| = 1$, hard case

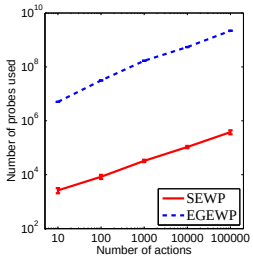

Figure 5. $|p| = \sqrt{K}$, easy case

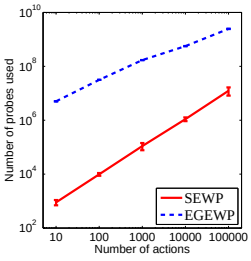

Figure 6. $|p| = \sqrt{K}$, hard case

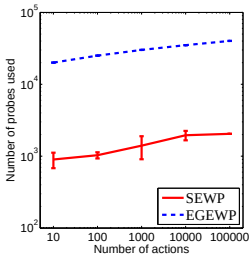

Figure 7. $|p| = K$, easy case

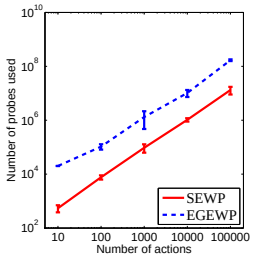

Figure 8. $|p| = K$, hard case

C.2. Experiment Settings for Figure 1

In Figure 1, the *lilUCB* algorithm comes from Jamieson et al. (2014). The parameters we used in experiments is the *lilUCB Heuristic* setting, which performs the best in the experiments of Jamieson et al. (2014). The *SE* algorithm is short for the *successive elimination* algorithm of Even-Dar et al. (2002). As these algorithms select options for measurements, we adapt them to the probe setting by choosing the first probe in some arbitrary ordering of probes that gives a measurement for the selected option . In experiments, all distributions we used are Gaussian with variance $1/4$. Each point reported in the figure is based on 100 repeated experiments under the same reward distributions, where we set one of the means to be 0.5 and the others to be 0.