Overlapping Multi-Bandit Best Arm Identification

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher

Abstract

In the multi-armed bandit literature, the multi-bandit best-arm identification problem consists of determining each best arm in a number of disjoint groups of arms, with as few total arm pulls as possible. In this paper, we introduce a variant of the multi-bandit problem with overlapping groups, and present two algorithms for this problem based on successive elimination and lower/upper confidence bounds (LUCB). We bound the number of total arm pulls required for high-probability best-arm identification in every group, and we complement these bounds with a near-matching algorithm-independent lower bound. In addition, we show that a specific choice of the groups recovers the top-k ranking problem.

I. Introduction

The multi-armed bandit (MAB) problem [1] provides a versatile framework for sequentially searching for high-reward actions, with applications including clinical trials [2], online advertising [3], adaptive routing [4], and portfolio design [5].

A variation of the MAB problem known as *multi-bandit best-arm identification* consists of finding the best arm in each of a number of separate groups of arms, while pulling the minimal *total* number of arms possible [6]. As a motivating example, consider a scenario where each arm corresponds to a product, and pulling an arm corresponds to testing how much it is liked by some user(s). Then the multi-bandit problem corresponds to searching for the top products among multiple separate types (e.g., TV, phone, music player, etc.).

Consider a variation of this example in which we not only want to find the top product of each type, but also the top products among several *overlapping* categories, e.g., top product under $\$100$, top product from each brand name, top newly-released product, and so on. This motivates the *overlapping multi-bandit best arm identification problem* (or overlapping multi-bandit problem for short), which we introduce and study in this paper. In a nutshell, we seek to find each best arm in a number of overlapping groups using as few total arm pulls as possible; see Section II for a formal description. Beyond the preceding example, the consideration of overlapping groups is of considerable interest when arms correspond to users, since categories such as gender, age, marital status, etc. invariably exhibit overlap.

A. Related Work

The literature on theory and algorithms for MAB problems is extensive; see [1], [7] for recent overviews. One of the main defining features of such problems is the distinction between stochastic vs. adversarial rewards; this paper focuses exclusively on the former.

Starting with early works such as [8], particular attention has been paid to *cumulative regret* measures. In contrast, this paper is more closely related to best arm identification, which has been solved using elimination methods [9]–[11], upper confidence bound (UCB) algorithms [12], [13], and lower/upper confidence bound (LUCB) algorithms [14], often with near-matching lower bounds [10], [12], [15]. A survey comparing these algorithmic approaches is given in [16].

A closely related problem is top-k identification [11], [14], for which recent developments have included near-tight bounds via successive elimination [17] and an LUCB-type algorithm with similar theoretical guarantees [18]. There is also a growing literature on active top-k ranking [19], [20] (i.e., not only identifying the top k arms, but also their order). In this setting, the emphasis has predominantly been on pairwise comparisons rather than regular bandit arm pulls, with a recent exception being [21].

To our knowledge, the first regret bounds for the multi-bandit problem were given in [6], adopting a "gap-based exploration" approach based on confidence bounds. A similar bound was obtained via a much simpler analysis using successive elimination [11], which also has the additional advantage of being parameter-free (while [6] requires knowledge of a certain complexity parameter).

B. Contributions

We introduce a novel variant of the multi-bandit problem with overlapping groups, provide two algorithms for solving this problem with rigorous guarantees upper bounding the number of arms pulled, and give a near-matching lower bound. Specifically, we first consider a simple successive elimination algorithm, and then a variant of LUCB [14] adapted to our setting.

Our setting trivially captures the regular multi-bandit problem, for which we recover similar results to those of [6], [11], as well a near-matching lower bound. In addition, we show that the top-k ranking problem with regular bandit feedback is a special case of our framework, and discuss the connection between this special case and the coarse ranking framework of [21].

II. Problem Setup

We consider a MAB setting with n arms having reward distributions $(\nu_1, \dots, \nu_n)$, the corresponding means of which are $(\mu_1, \dots, \mu_n)$ with $\mu_j \in (0,1)$. It is assumed that each ν_j is sub-Gaussian[1] with parameter $\sigma \leq \frac{1}{2}$; as noted in [16], this accounts for all distributions whose support is a subset of $[0,1]$ (e.g., Bernoulli).

As indicated above, the key novelty of our setting is allowing for general possibly-overlapping groups. Specifically, there is a known set of groups $\mathcal{G}$, where each $G \in \mathcal{G}$ is a subset of $\{1, \dots, n\}$. The number of groups is denoted by m, and the groups are denoted by $G_1, \dots, G_m$.

An algorithm for the overlapping multi-bandit problem iteratively pulls arms at times indexed by $t = 1, 2$, etc. At each time, the algorithm chooses an arm j_t and observes an independent reward $X_{j_t, T_j(t)} \sim \nu_{j_t}$, where $T_j(t)$ is the number of pulls of arm j up to time t. The empirical estimate of μ_j after $T_j(t)$ pulls of arm j is denoted by $\hat{\mu}_{j,T_j(t)} = \frac{1}{T_j(t)} \sum_{s=1}^{T_j(t)} X_{j,s}$.

At any given time, the algorithm may choose to stop and output m recommendations $\widehat{j}(G_1), \dots, \widehat{j}(G_m)$ as estimates of the best arms in the groups. The time at which this occurs is called the *stopping time*, and we would like it to be as small as possible. In addition, we seek the correct identification of the best arm in each group $G \in \mathcal{G}$. Writing the true best arm of group G (which is assumed to be unique) as

$$j^*(G) = \arg\max_{j \in G} \mu_j, \tag{1}$$

the error probability is given by

$$P_{\mathrm{e}} = \mathbb{P}\Big[\bigcup_{G \in \mathcal{G}} \big\{ \widehat{j}(G) \neq j^*(G) \big\} \Big]. \tag{2}$$

We are interested in algorithms that achieve $P_{\mathrm{e}} \leq \delta$ with guarantees on the total number of arm pulls (i.e., the stopping time), henceforth denoted by T.

Stochastic MAB problems invariably contain fundamental "gaps" between certain arms that dictate the required number of arm pulls. In our setting, these gaps are defined as follows:

$$\Delta_j = \min \left\{ \min_{G\,:\, j \in G, j = j^*(G)} \big(\mu_j - \mu_{j_{\mathrm{sec}}(G)}\big), \min_{G\,:\, j \in G, j \neq j^*(G)} \big(\mu_{j^*(G)} - \mu_j\big) \right\} \geq 0, \tag{3}$$

where $j_{\mathrm{sec}}(G)$ is the second-best arm in G, and the minimum of an empty set is infinity. Without loss of generality, we assume that each arm is in at least one group, and that all groups contain at least two arms; this implies that each Δ_j is well-defined and finite.

The assumption that $j^*(G)$ is uniquely defined in (1) is equivalent to requiring $\Delta_j > 0$ for all $j = 1, \dots, n$. We henceforth refer to any such instance as *identifiable*, and to all other instances as *non-identifiable*. We assume identifiability in all of our main results.

A. Auxiliary Results

Here we review some useful auxiliary results from the MAB literature that we will use in our analysis.

Early works on multi-armed bandits relied on basic concentration bounds such as Hoeffding's inequality to establish that the empirical mean of an arm approaches the true mean as it is pulled more. To improve certain logarithmic factors in the final results, we adopt a more recent approach based on the law of iterated logarithm [22].

Lemma 1. (Law of iterated logarithm [16, Lemma 1]) *Let $Z_1, Z_2, \dots$ be i.i.d. sub-Gaussian random variables with mean $\mu \in \mathbb{R}$ and parameter $\sigma \leq \frac{1}{2}$. For any $\epsilon \in (0,1)$ and $\delta \in \big(0, \frac{1}{e}\log(1+\epsilon)\big)$, it holds with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$ that*

$$\left| \frac{1}{t} \sum_{s=1}^{t} Z_s - \mu \right| \leq U(t, \delta), \quad \forall t \geq 1, \tag{4}$$

where

$$U(t,\delta) = (1 + \sqrt{\epsilon}) \sqrt{\frac{1+\epsilon}{2t} \log \frac{\log(1+\epsilon)t}{\delta}}. \tag{5}$$

In accordance with this result, we define the following *upper and lower confidence bounds* at time t:

$$\mathrm{UCB}_t(j) = \hat{\mu}_{j,T_j(t)} + U(T_j(t), \delta/n) \tag{6}$$

$$\mathrm{LCB}_t(j) = \hat{\mu}_{j,T_j(t)} - U(T_j(t), \delta/n), \tag{7}$$

where the division of δ by n is in accordance with a union bound over the n arms.

[1] A zero-mean random variable Z is sub-Gaussian with parameter σ if $\mathbb{E}[e^{\lambda Z}] \leq \exp\big(\frac{\lambda^2 \sigma^2}{2}\big)$. For a random variable Z with a non-zero mean, we use the terminology sub-Gaussian to mean that this property holds for $Z - \mathbb{E}[Z]$.

Corollary 1. (Confidence bounds) *If the arm reward distributions satisfy the conditions of Lemma 1, then for any $\epsilon \in (0,1)$ and $\delta \in \big(0, \frac{1}{e}\log(1+\epsilon)\big)$, it holds with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$ that*

$$\mathrm{LCB}_t(j) \le \mu_j \le \mathrm{UCB}_t(j), \quad \forall j \in \{1,\dots,n\}, t \ge 1. \tag{8}$$

Proof. This follows by applying Lemma 1 for each $j = 1,\dots,n$ with $Z_s = X_{j,s}$ and δ/n in place of δ, and taking the union bound over j. Note that after the union bound, $n \cdot \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta/n}{\log(1+\epsilon)}\big)^{1+\epsilon} \le \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$. □

In the analysis of the algorithms, we will need to "invert" $U(t,\delta)$ in the sense of establishing how large t needs to be to upper bound it by a certain threshold. Such an inversion is given as follows.

Lemma 2. (Inversion of $U(t,\delta)$ [16, Eq. (4)]) *The quantity $U(t,\delta)$ defined in* (5) *is such that, for any positive numbers (δ, n, Δ) with $\Delta \in (0,1)$, we have*

$$\min\Big\{k \,:\, U(k,\delta/n) \le \frac{\Delta}{4}\Big\} \le \frac{2\gamma}{\Delta^2}\log\frac{2\log\big(\gamma(1+\epsilon)\Delta^{-2}\big)}{\delta/n}, \tag{9}$$

where $\gamma = 8(1+\sqrt{\epsilon})^2(1+\epsilon)$.

Finally, the following lemma relating the number of arm pulls of two different instances permits a simple and elegant approach to establishing lower bounds. Here and subsequently, we let N_j denote the total number of times arm j has been pulled upon termination, so that $T = \sum_{j=1}^n N_j$.

Lemma 3. (Relating two instances [15, Lemma 1]) *Let $\nu = (\nu_1,\dots,\nu_n)$ and $\nu' = (\nu'_1,\dots,\nu'_n)$ be two different bandit instances such that for all $j = 1,\dots,n$, the distributions ν_j and ν'_j are mutually absolutely continuous. For any almost-surely finite stopping time σ, and any event $\mathcal{A}$ depending only on the history up to the stopping time, we have*

$$\sum_{j=1}^n \mathbb{E}_\nu[N_j(\sigma)] D(\nu_j \| \nu'_j) \ge d(\mathbb{P}_\nu[\mathcal{A}], \mathbb{P}_{\nu'}[\mathcal{A}]), \tag{10}$$

where $D(\nu_j\|\nu'_j) = \mathbb{E}_{\nu_j}\big[\log\frac{\nu_j(X)}{\nu'_j(X)}\big]$ is the KL divergence, and $d(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{a-b}$. In particular, if $\mathbb{P}_\nu[\mathcal{A}] \ge 1-\delta$ and $\mathbb{P}_{\nu'}[\mathcal{A}] \le \delta$ for some $\delta \in (0,1)$, then[2]

$$\sum_{j=1}^n \mathbb{E}_\nu[N_j(\sigma)] D(\nu_j\|\nu'_j) \ge \log\frac{1}{2.4\delta}. \tag{11}$$

III. Lower Bound

In this section, we establish a performance benchmark for our practical algorithms by providing an algorithm-independent lower bound on the average number of arm pulls when $P_{\mathrm{e}} \le \delta$.

We assume in this section that the MAB reward distributions $(\nu_1,\dots,\nu_n)$ satisfy the following assumption.

Assumption 1. Each distribution ν_j in the bandit instance $(\nu_1,\dots,\nu_n)$ comes from a parametric family $\mathcal{P}$, and is uniquely parametrized by its mean $\mu_j \in (0,1)$. In addition, any two distributions $\nu_j, \nu'_j \in \mathcal{P}$ are mutually absolutely continuous, and $D(\nu_j\|\nu'_j) \to 0$ as the means of ν_j and ν'_j approach each other.

Assumption 1 is satisfied for Bernoulli rewards and Gaussian rewards with a fixed variance, among others [8], [15]. Our analysis can also readily be extended to more general families of arm distributions satisfying [15, Assumption 3], which essentially states that any given arm distribution can be replaced by a different distribution with a strictly higher (or strictly smaller) mean but a similar KL divergence to a reference arm (see also [8]).

Our first main result is given as follows.

Theorem 1. (Lower bound) *Under Assumption 1, suppose that a given algorithm $\mathtt{Alg}^*$ achieves $P_{\mathrm{e}} \le \delta$ for all identifiable bandit instances with reward distributions in $\mathcal{P}$. Fix an identifiable instance $(\nu_1,\dots,\nu_n)$ with means $(\mu_1,\dots,\mu_n)$, and for each $j = 1,\dots,n$, let $\nu'_j \in \mathcal{P}$ be defined via its mean μ'_j as follows for sufficiently small $\alpha > 0$:*[3]

- *If the outer minimum in* (3) *is achieved by the first term (i.e., by a group in which j is best) then $\mu'_j = \mu_j - (1+\alpha)\Delta_j$;*
- *Otherwise, $\mu'_j = \mu_j + (1+\alpha)\Delta_j$.*

[2] See [15, Remark 2] for this variation.

[3] Specifically, $\alpha > 0$ is arbitrary subject to being sufficiently small so that $\mu'_j \in (0,1)$. This is possible due to the fact that $\mu_j \in (0,1)$.

When $\texttt{Alg}^*$ *is run on the instance* $(\nu_1, \dots, \nu_n)$*, the average number of arm pulls is at least* $T_{\mathrm{lower}}(\delta)$*, where*

$$T_{\mathrm{lower}}(\delta) = \sum_{j=1}^{n} \frac{\log \frac{1}{2.4\delta}}{D(\nu_j \| \nu_j')}. \tag{12}$$

Proof. Fix a given arm j, and let $\nu^{(j)}$ be the instance where ν_j is replaced by ν_j', and all other arms remain the same as ν. We observe from the definition of ν_j' in the theorem statement that this change alters one group's best arm. In the first case, there is a group where j was best but it is pushed below the second-best, and in the second case, there is a group where j was not best but it is pushed above the best. Note that the definition of ν_j' via its mean μ_j' is valid due to Assumption 1, and the mutual absolute continuity condition therein ensures that $D(\nu_j \| \nu_j')$ is finite.

In the following, we assume that $\nu^{(j)}$ is also an identifiable instance, i.e., each group has a unique best arm. In the supplementary material, we provide the required changes to circumvent this assumption; these changes make use of the final part of Assumption 1.

Letting $\mathcal{A}$ in (10) be the event that the algorithm provides the correct output for ν (and hence, an incorrect output for $\nu^{(j)}$), we claim that Lemma 3 yields

$$\mathbb{E}_\nu[N_j] \geq \frac{\log \frac{1}{2.4\delta}}{D(\nu_j \| \nu_j')}. \tag{13}$$

Indeed, this follows from the fact that $P_{\mathrm{e}} \leq \delta$ on all identifiable instances, and since by construction the KL divergence for arms indexed by $j' \neq j$ is zero (i.e., the distributions are identical in the two instances).

Since (13) holds for any j, the average number of arm pulls is lower bounded by the sum of the right-hand side over all j, thus proving (12). □

Remark 1. The bound (12) takes the same form as our upper bounds (to be given in the subsequent sections) whenever $D(\nu_j \| \nu_j') \leq c\Delta_j^2$ for some constant c, in which case

$$T_{\mathrm{lower}}(\delta) \geq \sum_{j=1}^{n} \frac{\log \frac{1}{2.4\delta}}{c\Delta_j^2}. \tag{14}$$

For instance, under Gaussian rewards with variance σ^2, a standard calculation gives $D(\nu_j \| \nu_j') = \frac{\Delta_j^2(1+\alpha)^2}{2\sigma^2}$. Moreover, under Bernoulli rewards with means in the range $(\eta, 1-\eta)$, it is known that $D(\nu_j \| \nu_j') \leq \frac{\Delta_j^2(1+\alpha)^2}{\eta(1-\eta)}$ [7, Eq. (2.8)]. See [8, Sec. 4] for KL divergence calculations for other families of arm distributions.

IV. Successive Elimination Algorithm

Successive elimination is a common MAB technique in which confidence bounds are used to rule out suboptimal arms, the remaining arms are sampled once each, and this procedure is repeated until one arm remains. In this section, we adopt this approach for the overlapping multi-bandit problem.

As is common in elimination algorithms, we work in *epochs* indexed by $i = 1, 2, \dots$, where within a given epoch we pull several arms. To decide which arms to pull and which to eliminate, we make use of the following definitions:

- Potential maximizers within group G. This is the set of arms $j \in G$ whose UCB is at least as high as the highest LCB:

$$M_i^{(G)} = \left\{ j \in G \,:\, \mathrm{UCB}_{t_i}(j) \geq \max_{j' \in G} \mathrm{LCB}_{t_i}(j') \right\} \tag{15}$$

  under the definitions (6)–(7), where t_i is the total number of arm pulls after those that occur in the i-th epoch.
- Unresolved groups. This is the set of groups that still have at least two potential maximizers:

$$\widetilde{\mathcal{G}}_i = \big\{ G \in \mathcal{G} \,:\, |M_i^{(G)}| \geq 2 \big\}, \tag{16}$$

  with $\widetilde{\mathcal{G}}_0 = \mathcal{G}$.
- Arms of interest. This is the set of arms that are the potential maximizer for at least one unresolved group:

$$\mathcal{A}_i = \big\{ j \,:\, \exists G \in \widetilde{\mathcal{G}}_i \text{ with } j \in G \big\}. \tag{17}$$

With these definitions in place, the successive elimination algorithm is described in Algorithm 1.

Theorem 2. (Upper bound for successive elimination) *For any* $\epsilon \in (0,1)$ *and* $\delta \in \big(0, \frac{1}{e}\log(1+\epsilon)\big)$*, with probability at least* $1 - \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$*, the successive elimination algorithm terminates with the correct output after at most* $T_{\mathrm{elim}}(\delta, \epsilon)$ *arm pulls, where*

$$T_{\mathrm{elim}}(\delta, \epsilon) = \sum_{j=1}^{n} \frac{2\gamma}{\Delta_j^2} \log \frac{2\log\big(\gamma(1+\epsilon)\Delta_j^{-2}\big)}{\delta/n}, \tag{18}$$

Algorithm 1 Successive Elimination Algorithm for the Overlapping Multi-Bandit Problem

Require: Groups $\mathcal{G}$, constants $\delta, \epsilon > 0$

1: Initialize $i = 1$, $t = 0$, and $T_j(t) = 0$ $(\forall j)$
2: Set $M_0^{(G)} = G$ $(\forall G)$, $\widetilde{\mathcal{G}}_0 = \mathcal{G}$, and $\mathcal{A}_0 = \{1, \dots, n\}$
3: **while** $\mathcal{A}_{i-1} \neq \emptyset$ **do**
4: Pull every arm in $\mathcal{A}_{i-1}$ once, incrementing t after each pull and updating all $T_j(t)$
5: Compute $M_i^{(G)}$, $\widetilde{\mathcal{G}}_i$ and $\mathcal{A}_i$ via (15)–(17)
6: For all G with $|M_i^{(G)}| = 1$, set $\widehat{j}(G)$ to be the corresponding single arm.
7: Increment the epoch index i
8: **end while**
9: **return** $(\widehat{j}(G_1), \dots, \widehat{j}(G_m))$

with $\gamma = 8(1 + \sqrt{\epsilon})^2(1 + \epsilon)$.

Proof. It suffices to show that when the high-probability event in Corollary 1 holds, the algorithm terminates with the correct estimates and performs at most $T_{\mathrm{elim}}(\delta, \epsilon)$ arm pulls.

We first show that the algorithm never removes an optimal arm $j^*(G)$ from the arms of interest without first correctly assigning $\widehat{j}(G) = j^*(G)$ (for all G in which it is optimal). We prove this by induction, with the trivial base case being that $j^*(G)$ is initially both of interest and in $M_0^{(G)}$ by construction. Now, assuming $j^*(G) \in M_{i-1}^{(G)}$ after the $(i-1)$-th epoch, we have

$$\begin{aligned} \mathrm{UCB}_{t_i}(j^*(G)) &\geq \mu_{j^*(G)} && (19)\\ &= \max_{j' \in G} \mu_{j'} && (20)\\ &\geq \max_{j' \in G} \mathrm{LCB}_{t_i}(j'), && (21) \end{aligned}$$

where both (19) and (21) use the validity of the confidence bounds (Corollary 1). Therefore, $j^*(G)$ meets the condition (15), and it remains in $M_i^{(G)}$ in Line 6 of Algorithm 1. By induction, this means that $j^*(G)$ remains in $M_i^{(G)}$ as long as G remains unresolved, so it is the only arm in G that can be declared optimal.

Next, we bound the number of pulls of each arm. By construction, after Line 4 of Algorithm 1 in a given epoch, all arms of interest have been pulled the same number of times, and therefore have the same value of $U(T_j(t), \delta/n)$, henceforth referred to as U_i in epoch i. By Lemma 2, this value is at most $\frac{\Delta}{4}$ once the number of epochs reaches the right-hand side of (9). Now, fix any arm j and note the following:

- If j is the top arm in group G, the group will be resolved once all other $j' \neq j$ from G are removed from $M_i^{(G)}$. For any such j', if $U_i < \frac{\Delta_j}{4}$ then by (6)–(7), $|\mathrm{UCB}_{t_i}(j') - \mathrm{LCB}_{t_i}(j')| < \frac{\Delta_j}{2}$. Hence,

$$\begin{aligned} \mathrm{UCB}_{t_i}(j') &< \mathrm{LCB}_{t_i}(j') + \frac{\Delta_j}{2} && (22)\\ &\leq \mu_{j'} + \frac{\Delta_j}{2} && (23)\\ &\leq \mu_j - \frac{\Delta_j}{2} && (24)\\ &\leq \mathrm{UCB}_{t_i}(j) - \frac{\Delta_j}{2} && (25)\\ &< \mathrm{LCB}_{t_i}(j), && (26) \end{aligned}$$

  where (22) and (26) use the above-mentioned gap between UCB and LCB, (23) and (25) use the validity of the confidence bounds, and (24) uses the definition of Δ_j and the fact that $j = j^*(G)$.

  We see from (26) that j' is removed from $M_i^{(G)}$. Since this holds for all $j' \neq j$ in G, it follows that the group is marked as resolved.

Algorithm 2 LUCB Algorithm for the Overlapping Multi-Bandit Problem

Require: Groups $\mathcal{G}$, constants $\delta, \epsilon > 0$

Sample each arm once; set $T_j(n) \leftarrow 1$ $(\forall j)$; initialize $t = n$ and $i = 1$
while True **do**
  for $G \in \mathcal{G}$ **do**
    $h_i^{(G)} = \arg\max_{j \in G} \ \hat{\mu}_{j,T_j(t)}$
    $l_i^{(G)} = \arg\max_{j \in G \setminus \{h_i^{(G)}\}} \ \mathrm{UCB}_t(j)$
    $w_i^{(G)} = \mathrm{UCB}_t(l_i^{(G)}) - \mathrm{LCB}_t(h_i^{(G)})$
  end for
  $G_i' \leftarrow \arg\max_{G \in \mathcal{G}} w_i^{(G)}$ (breaking ties arbitrarily)
  if $w_i^{(G_i')} \leq 0$ **then**
    return $(h_i^{(G_1)}, \ldots, h_i^{(G_m)})$
  else
    Sample $h_i^{(G_i')}$ and $l_i^{(G_i')}$
    Set $t \leftarrow t + 2$ and $i \leftarrow i + 1$; update all $T_j(t)$
  end if
end while

- On the other hand, if $j \in G$ is not the top arm in G, and if $U_i < \frac{\Delta_j}{4}$, then

$$\mathrm{UCB}_{t_i}(j) < \mathrm{LCB}_{t_i}(j) + \frac{\Delta_j}{2} \tag{27}$$
$$\leq \mu_j + \frac{\Delta_j}{2} \tag{28}$$
$$\leq \mu_{j^*(G)} - \frac{\Delta_j}{2} \tag{29}$$
$$\leq \mathrm{UCB}_{t_i}(j^*(G)) - \frac{\Delta_j}{2} \tag{30}$$
$$< \mathrm{LCB}_{t_i}(j^*(G)), \tag{31}$$

  by the same arguments as (22)–(26). We see that (31) implies that j is removed from $M_i^{(G)}$.

Combining these cases, we conclude that arm j only ever continues being pulled if $U_i \geq \frac{\Delta_j}{4}$. Since i is precisely the number of arm pulls of all remaining arms after epoch i, applying Lemma 2 and summing (9) over $j = 1, \ldots, n$ yields (18). □

We observe that (18) matches (14) up to the constant factors and the extra log factor $\log \frac{2\log(\gamma(1+\epsilon)\Delta_j^{-2})}{n}$, which is typically insignificant compared to the leading $\frac{1}{\Delta_j^2}$ term.

In particular, if $\delta = O(n^{-\alpha})$ and $\min_j \Delta_j = \Omega(n^{-\beta})$ for some constants $\alpha, \beta > 0$, then we find that the factors $\log \frac{2\log\left(\gamma(1+\epsilon)\Delta_j^{-2}\right)}{\delta/n}$ and $\log \frac{1}{2.4\delta}$ both simplify to $O(\log n)$. Hence, in this case, the upper and lower bounds match to within a constant factor.

V. LUCB-Type Algorithm

In Algorithm 2, we describe a lower-upper confidence bound (LUCB) algorithm inspired by that proposed for top-k identification [14], [16]. We initially pull every arm once, and then proceed in rounds within which two arms are pulled; similarly to Algorithm 1, these rounds are indexed by $i \geq 1$.

In round i, within each group $G \in \mathcal{G}$, we consider the highest-mean arm $h_i^{(G)}$, and the arm $l_i^{(G)}$ with the highest UCB score in $G \setminus \{h_i^{(G)}\}$. If $\mathrm{UCB}_t(l_i^{(G)}) - \mathrm{LCB}_t(h_i^{(G)}) < 0$ for all $G \in \mathcal{G}$, then we believe each $h_i^{(G)}$ to be optimal within its group, so we terminate. Otherwise, to learn more about the competing arms $h_i^{(G)}$ and $l_i^{(G)}$, we pull them both for the group such that their confidence regions overlap the most (i.e., $\mathrm{UCB}_t(l_i^{(G)}) - \mathrm{LCB}_t(h_i^{(G)})$ is highest). As usual, here t denotes the total number of arm pulls so far.

Theorem 3. (Upper bound for LUCB) *For any $\epsilon \in (0,1)$ and $\delta \in \left(0, \frac{1}{e}\log(1+\epsilon)\right)$, with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}$, the LUCB algorithm terminates with the correct output after at most $T_{\mathrm{lucb}}(\delta,\epsilon)$ arm pulls, where*

$$T_{\mathrm{lucb}}(\delta,\epsilon) = 2\sum_{j=1}^{n} \frac{2\gamma}{\Delta_j^2} \log \frac{2\log\left(\gamma(1+\epsilon)\Delta_j^{-2}\right)}{\delta/n}, \tag{32}$$

with $\gamma = 8(1+\sqrt{\epsilon})^2(1+\epsilon)$.

Proof. We first show that when the high probability event in Corollary 1 holds, the algorithm can only terminate with the correct output $(j^*(G_1), \ldots, j^*(G_m))$. Suppose for the purpose of contradiction that the algorithm terminates during round i and returns $(h_i^{(G_1)}, \ldots, h_i^{(G_m)}) \neq (j^*(G_1), \ldots, j^*(G_m))$. This implies that there is at least one group G for which $h_i^{(G)} \neq j^*(G)$. Letting t_i denote the time index in Line 6 of Algorithm 2 during round i, we have

$$\begin{aligned}
\mu_{h_i^{(G)}} &\geq \mathrm{LCB}_{t_i}(h_i^{(G)}) && (33)\\
&\geq \mathrm{UCB}_{t_i}(l_i^{(G)}) && (34)\\
&\geq \mathrm{UCB}_{t_i}(j^*(G)) && (35)\\
&\geq \mu_{j^*(G)}, && (36)
\end{aligned}$$

where (33) and (36) use the validity of the confidence bounds (Corollary 1), (34) uses the stopping condition, and (35) uses the definition of $l_i^{(G)}$. From (36), we have $\mu_{h_i^{(G)}} \geq \mu_{j^*(G)}$, which is in contradiction with $j^*(G)$ being the unique best arm in G. Hence, under the event in Corollary 1, the algorithm will never return the wrong output.

Next, we bound the number of pulls of each arm. Define $c(G) := \frac{\mu_{j^*(G)} + \mu_{j_{\mathrm{sec}}(G)}}{2}$, where $\mu_{j_{\mathrm{sec}}(G)}$ is the second best arm in group G. We say that an arm $j \in G$ is *G-BAD* for the group G in round i if either of the following two conditions hold:

$$\begin{aligned}
&j = j^*(G) \text{ and } \mathrm{LCB}_{t_i}(j) < c(G), \text{ or} && (37)\\
&j \neq j^*(G) \text{ and } \mathrm{UCB}_{t_i}(j) > c(G). && (38)
\end{aligned}$$

Recall that G_i' is the group from which $h_i^{(G_i')}$ and $l_i^{(G_i')}$ are selected in round i. For all $i \geq 1$, conditioned on the event in Corollary 1, we claim that

$$\begin{aligned}
&\mathrm{LCB}_{t_i}(h_i(G_i'))) < \mathrm{UCB}_{t_i}(l_i(G_i'))) \implies\\
&\{h_i(G_i') \text{ is } G_i'\text{-BAD}\} \text{ or } \{l_i(G_i') \text{ is } G_i'\text{-BAD}\}. && (39)
\end{aligned}$$

Simply put, (39) states that if the stopping condition is not satisfied then either $h_i(G_i')$ or $l_i(G_i')$ is G_i'-*BAD*. We prove (39) in the supplementary material.

For an arm j, let τ_j denote the smallest integer such that $U(\tau_j, \delta/n) \leq \frac{\Delta_j}{4}$. We show that if j has been pulled some number of times $q \geq \tau_j$ in a given round, then j cannot be *G-BAD* for any group G containing j. We proceed by considering the following two cases:

- **Case** $j \notin \{j^*(G_1), \ldots, j^*(G_m)\}$:
  In this case, we have
  $$\Delta_j = \min_{G\,:\,j \in G} \mu_{j^*(G)} - \mu_j. \tag{40}$$
  Supposing $q \geq \tau_j$, let G be any group such that $j \in G$. Then, we have
  $$\begin{aligned}
  \hat{\mu}_{j,q} + U(q, \delta/n) &\leq \mu_j + 2U(q, \delta/n) && (41)\\
  &= \frac{\mu_{j^*(G)} + \mu_{j_{\mathrm{sec}}(G)}}{2} + 2U(q, \delta/n)\\
  &\quad + \frac{(\mu_j - \mu_{j^*(G)}) + (\mu_j - \mu_{j_{\mathrm{sec}}(G)})}{2} && (42)\\
  &\leq c(G) + 2U(q, \delta/n) + \frac{\mu_j - \mu_{j^*}(G)}{2} && (43)\\
  &\leq c(G) + \frac{\Delta_j}{2} + \frac{\mu_j - \min_{G:j\in G} \mu_{j^*(G)}}{2} && (44)\\
  &= c(G) + \frac{\Delta_j}{2} - \frac{\Delta_j}{2} = c(G), && (45)
  \end{aligned}$$
  where (41) uses the validity of the confidence bounds, (43) follows from $\mu_j - \mu_{j_{\mathrm{sec}}(G)} \leq 0$ and the definition of $c(G)$, (44) uses $U(q, \delta/n) \leq \frac{\Delta_j}{4}$, and (45) uses (40). This result implies that j cannot be *G-BAD* for any group G containing j.
- **Case** $j \in \{j^*(G_1), \ldots, j^*(G_m)\}$:
  We consider the following two sub-cases:

– For $G \in \{G : j \in G \text{ and } j = j^*(G)\}$, we have

$$\hat{\mu}_{j,q} - U(q, \delta/n) \geq \mu_j - 2U(q, \delta/n) \tag{46}$$

$$= \frac{\mu_{j^*(G)} + \mu_{j_{\sec}(G)}}{2} - 2U(q, \delta/n) + \frac{\mu_{j^*(G)} - \mu_{j_{\sec}(G)}}{2} \tag{47}$$

$$\geq c(G) - \frac{\Delta_j}{2} + \frac{\mu_{j^*(G)} - \mu_{j_{\sec}(G)}}{2} \tag{48}$$

$$\geq c(G) - \frac{\Delta_j}{2} + \frac{\Delta_j}{2} = c(G) \ , \tag{49}$$

where (47) follows from $j = j^*(G)$, (48) from $U(q, \delta/n) \leq \frac{\Delta_j}{4}$, and (49) from the fact that $\mu_{j^*(G)} - \mu_{j_{\sec}(G)} \geq \Delta_j$ when $j = j^*(G)$. This implies that j cannot be *G-BAD* for any group $G \in \{G : j \in G \text{ and } j = j^*(G)\}$.

– For $G \in \{G : j \in G \text{ and } j \neq j^*(G)\}$, we have

$$\hat{\mu}_{j,q} + U(q, \delta/n) \leq \mu_j + 2U(q, \delta/n) \tag{50}$$

$$= \frac{\mu_{j^*(G)} + \mu_{j_{\sec}(G)}}{2} + 2U(q, \delta/n) - \frac{(\mu_{j^*(G)} - \mu_j) + (\mu_{j_{\sec}(G)} - \mu_j)}{2} \tag{51}$$

$$\leq c(G) + \frac{\Delta_j}{2} - \frac{\mu_{j^*(G)} - \mu_j}{2} \tag{52}$$

$$\leq c(G) + \frac{\Delta_j}{2} - \frac{\Delta_j}{2} = c(G) \ , \tag{53}$$

where (52) follows from $U(q, \delta/n) \leq \frac{\Delta_j}{4}$ and $\mu_{j_{\sec}(G)} - \mu_j \geq 0$, and (53) from the fact that $\mu_{j^*(G)} - \mu_j \geq \Delta_j$ when $j \neq j^*(G)$. This implies that j cannot be *G-BAD* for any group $G \in \{G : j \in G \text{ and } j \neq j^*(G)\}$.

In summary, the results in these sub-cases imply that $j \in \{j^*(G_1), \ldots, j^*(G_m)\}$ cannot be *G-BAD* for any group G containing j if j has been sampled τ_j or more times and $j \in \{j^*(G_1), \ldots, j^*(G_m)\}$.

Combining the above, we deduce that conditioned on the high probability event from Corollary 1, the total number of rounds does not exceed the following:

$$\sum_{i=1}^{\infty} \mathbf{1}\{h_i^{(G_i')} \text{ is } G_i'\text{-BAD or } l_i^{(G_i')} \text{ is } G_i'\text{-BAD}\} \tag{54}$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{n} \mathbf{1}\{\{h_i^{(G_i')} = j \text{ or } l_i^{(G_i')} = j\} \cap \{j \text{ is } G_i'\text{-BAD}\}\} \tag{55}$$

$$\leq \sum_{i=1}^{\infty} \sum_{j=1}^{n} \mathbf{1}\{\{h_i^{(G_i')} = j \text{ or } l_i^{(G_i')} = j\} \cap \{T_j(t_i) < \tau_j\}\} \tag{56}$$

$$\leq \sum_{j=1}^{n} (\tau_j - 1), \tag{57}$$

where (57) follows since $T_j(t_1) = 1$ by construction, and since $T_j(t_{i+1}) = T_j(t_i) + 1$ whenever $h_i^{(G_i')} = j$ or $l_i^{(G_i')} = j$. The proof of Theorem 3 is completed by noting that the total number of arm pulls is $n + (2 \times$ number of rounds), and substituting the upper bound in (9) (with $\Delta = \Delta_j$) for τ_j. □

Observe that the bounds in Theorems 2 and 3 coincide up to a factor of two, and hence, the guarantee of LUCB is also guaranteed to match the lower bound up to a logarithmic or constant factor.

VI. Application to Top-K Ranking

In the top-k ranking problem (e.g., see [19], [20]), we seek to identify the best k arms *and* their order with respect to the means $\{\mu_j\}_{j=1}^n$. This is in contrast with the more commonly considered top-k identification problem, where we do not care about the order [11], [14].

The following reduction shows that top-k ranking is a special case of our setting: Let the $m = \binom{n}{k-1}$ groups be all the subsets of $\{1, ..., n\}$ of size $n - k + 1$. This group size ensures that every group has at least one top-k arm, and yields the following:

1) An arm is in the set of top k arms if and only if it is the best arm in at least one group. Therefore, if we know the best arm in each group, then we also know which arms are the top k.
2) Given knowledge of the best arm in each group, the ordering within the top k is uniquely identified by observing, for each (i, j) in the top k, which arm is best in the group $\{i\} \cup \{j\} \cup \mathcal{A}_{ij}$, where $\mathcal{A}_{ij}$ is an arbitrary subset of size $n-k-1$ among the bottom $n-k$ arms (which are known due to the previous item).

Conversely, if the top-k ranking solution is known, we can trivially establish the best arm in each such group. We conclude that the top-k ranking problem and the overlapping multi-bandit problem are equivalent under the choice of groups described above.

With this reduction in place, we have the following corollary, where we let $\mu_{[1]}, \dots \mu_{[n]}$ denote a re-ordering of the arm means such that

$$\mu_{[1]} > \dots > \mu_{[k]} > \mu_{[k+1]} \geq \dots \geq \mu_{[n]}. \tag{58}$$

Here the strict inequalities ensure that the top-k ranking solution is uniquely defined.

Corollary 2. (Application to top-k ranking) *In the top-k ranking problem, the statements of Theorem 1, Theorem 2, and Theorem 3 hold true with the definition of Δ_j specialized as follows:*

$$\Delta_j = \begin{cases} \mu_{[1]} - \mu_{[2]} & j \text{ is best} \\ \min\{\mu_{[i]} - \mu_{[i+1]}, \mu_{[i-1]} - \mu_{[i]}\} & j \text{ is } i\text{-th best} \\ \mu_{[k]} - \mu_j & \text{otherwise,} \end{cases} \tag{59}$$

where the middle case holds for $2 \leq i \leq k$.

The equivalence of (3) and (59) is easily established via the above-mentioned group structure and reduction.

In the case of Bernoulli rewards, similar results to Corollary 2 can be deduced from the study of active coarse ranking in [21]. Moreover, the analysis therein can be extended to general sub-Gaussian rewards with relatively little difficulty. Thus, we do not claim the bounds in Corollary 2 themselves to have any significant novelty, but rather, our goal here is to highlight the strong connection between the two seemingly unrelated bandit settings.

At first glance, it may appear that implementing Algorithms 1 and 2 requires prohibitively large computation due to steps that search over $\binom{n}{k-1}$ groups. However, due to the structure of these groups, both algorithms can in fact be implemented efficiently via sorting; for Algorithm 2, this again produces a similar algorithm to the LUCB-type algorithm proposed in [21]. The details are given in the supplementary material.

VII. Conclusion

Motivated by overlapping group structures in practical multi-armed bandit applications, we have introduced and studied a novel overlapping multi-bandit best arm identification problem. Our algorithms based on successive elimination and LUCB-type selection are near-optimal, matching the lower bound up to a logarithmic factor in the general case, and up to a constant factor in broad scaling regimes on the error probability and gaps $\{\Delta_j\}_{j=1}^n$. In addition, we showed that our results apply directly to the problem of top-k ranking with regular bandit rewards, thus complementing the existing literature on top-k identification and top-k ranking via pairwise comparisons.

Appendix

A. Non-identifiable Instances in the Lower Bound

It may be the case that $\nu^{(j)}$ constructed in the proof of Theorem 1 is not an identifiable instances, as a result of μ_j being pushed below multiple $\mu_{j'}$ that were tied for second best in some group(s). In this case, we further modify $\nu^{(j)}$ so that in any groups with ties, one of the tied arms is shifted up by an arbitrarily small (i.e., essentially infinitesimal) amount to become the unique maximizer.

This modification leads to additional terms on the left-hand side of (10), but since each corresponding KL divergence can be made arbitrary small[4] by reducing the amount by which the shift is done above (*cf.*, final part of Assumption 1), all such terms can be neglected. Therefore, (13) still holds, and Theorem 1 remains true.

[4]The notion of "arbitrarily small" here can even be as a function of n and ν, so that this argument remains valid even when we sum over all the arms and incorporate the multiplications by $\mathbb{E}_\nu[N_{j'}(\sigma)]$ in (10)

B. Proof of G_i'-BAD Property (39)

Recall that G_i' is the group from which $h_i(G_i')$ and $l_i(G_i')$ are selected in round i. For all $i \geq 1$, conditioned on the event in Corollary 1, we prove that

$$\mathrm{LCB}_{t_i}(h_i(G_i')) < \mathrm{UCB}_{t_i}(l_i(G_i')) \implies \{h_i(G_i') \text{ is } G_i'\textit{-BAD}\} \text{ or } \{l_i(G_i') \text{ is } G_i'\textit{-BAD}\}.$$

Let τ denote the stopping round of the algorithm, i.e. the first round i such that $\mathrm{LCB}_{t_i}(h_i(G_i')) \geq \mathrm{UCB}_{t_i}(l_i(G_i'))$. Then the previous relation can be written as:

$$\{i < \tau\} \implies \{h_i(G_i') \text{ is } G_i'\textit{-BAD}\} \text{ or } \{l_i(G_i') \text{ is } G_i'\textit{-BAD}\}.$$

We prove this by contradiction by considering the following cases in which we assume that both $h_i(G_i')$ and $l_i(G_i')$ are not G_i'-BAD:

- **Case 1:** Using the stopping condition and the G_i'-BAD property, we have

$$\begin{aligned}
&\{i < \tau\} \text{ and } \{h_i(G_i') = j^*(G_i') \text{ is not } G_i'\text{-BAD}\} \text{ and } \{l_i(G_i') \neq j^*(G_i') \text{ is not } G_i'\text{-BAD}\} && (60)\\
&\implies \{\mathrm{LCB}_{t_i}(h_i(G_i')) < \mathrm{UCB}_{t_i}(l_i(G_i'))\} \text{ and } \{\mathrm{LCB}_{t_i}(h_i(G_i')) \geq c(G_i')\} \text{ and } \{\mathrm{UCB}_{t_i}(l_i(G_i')) \leq c(G_i')\} && (61)\\
&\implies \{\mathrm{LCB}_{t_i}(h_i(G_i')) < \mathrm{UCB}_{t_i}(l_i(G_i'))\} \text{ and } \{\mathrm{LCB}_{t_i}(h_i(G_i')) \geq \mathrm{UCB}_{t_i}(l_i(G_i'))\}. && (62)
\end{aligned}$$

  The obtained events are in contradiction.
- **Case 2:** Using the definition of the G_i'-BAD property, we have

$$\begin{aligned}
&\{i < \tau\} \text{ and } \{h_i(G_i') \neq j^*(G_i') \text{ is not } G_i'\text{-BAD}\} \text{ and } \{l_i(G_i') = j^*(G_i') \text{ is not } G_i'\text{-BAD}\} && (63)\\
&\implies \{\mathrm{UCB}_{t_i}(h_i(G_i')) \leq c(G_i')\} \text{ and } \{\mathrm{LCB}_{t_i}(l_i(G_i')) \geq c(G_i')\} && (64)\\
&\implies \{\hat{\mu}_{h_i(G_i'),T_{h_i(G_i')}(t_i)} < c(G_i')\} \text{ and } \{\hat{\mu}_{l_i(G_i'),T_{l_i(G_i')}(t_i)} > c(G_i')\}, && (65)
\end{aligned}$$

  where we have used the fact that $U(t,\delta/n)$ is always strictly positive in (6)–(7). It follows that $\hat{\mu}_{h_i(G_i'),T_{h_i(G_i')}(t_i)} < \hat{\mu}_{l_i(G_i'),T_{l_i(G_i')}(t_i)}$ which is in contradiction with $\hat{\mu}_{h_i(G_i'),T_{h_i(G_i')}(t_i)} = \arg\max_{j\in G_i'} \hat{\mu}_{j,T_j(t_i)}$.
- **Case 3:** Using the definition of the G_i'-BAD property, we have

$$\begin{aligned}
&\{i < \tau\} \text{ and } \{h_i(G_i') \neq j^*(G_i') \text{ is not } G_i'\text{-BAD}\} \text{ and } \{l_i(G_i') \neq j^*(G_i') \text{ is not } G_i'\text{-BAD}\} && (66)\\
&\implies \{\mathrm{UCB}_{t_i}(h_i(G_i')) \leq c(G_i')\} \text{ and } \{\mathrm{UCB}_{t_i}(l_i(G_i')) \leq c(G_i')\} && (67)\\
&\implies \{\mathrm{UCB}_{t_i}(j^*(G_i')) \leq c(G_i')\}. && (68)
\end{aligned}$$

  Since $j^*(G_i')$ is the unique best arm in G_i', we have $\mu_{j^*(G_i')} > c(G_i')$, which is in contradiction with the obtained event.

C. Efficient Implementations for Top-k Ranking

We showed in Section VI that we recover the top-k ranking problem upon letting the $m = \binom{n}{k-1}$ groups be all the subsets of $\{1,...,n\}$ of size $n-k+1$. Naively using this fact in Algorithms 1 and 2 leads to inefficient algorithms that iterate over all $\binom{n}{k-1}$ such subsets. However, here we show that both algorithms permit equivalent versions that have low computational complexity.

1) Successive Elimination: To describe the efficient implementation of successive elimination, we first present the following definitions with respect to the upper and lower confidence bounds in (6)–(7):

- If the LCB of arm j is above the UCB of arm j', then we say that j is *certifiably better* than j', and that j' is *certifiably worse* than j.
- We call an arm *certifiably i-th best* if it is certifiably worse than $i-1$ arms and certifiably better than $n-i$ arms.
- We call an arm *potentially top-k* if it is not certifiably worse than k or more other arms.
- We call an arm *of interest* if both of the following hold: (i) It is potentially top-k, (ii) It is not certifiably i-th best for any $i = 1,\dots,k$.

With these definitions, the algorithm proceeds as per Algorithm 1, pulling every arm of interest in a given epoch and then updating those arms of interest. Such updates can be done efficiently by sorting the relevant UCB and LCB scores.

To see the equivalence to Algorithm 1, we need to show that the two notions of "arm of interest" are identical. To see this, recall the definition of the potential maximizers $M_t^{(G)}$ in (15), and note the following for a given arm j:

- Suppose that conditions (i) and (ii) in the final dot point above hold. By the first condition, we know that there exists a set $\mathcal{A}_0$ of $n-k$ arms that j is potentially better than (i.e., not certifiably worse). Moreover, by the second condition, there exists an arm $j' \neq j$ (not necessarily in $\mathcal{A}_0$) that is neither certifiably better nor certifiably worse than j. If we consider a group G equaling a subset of $\mathcal{A}_0 \cup \{j\} \cup \{j'\}$ containing both j and j', then we immediately deduce that $\{j,j'\} \subseteq M_t^{(G)}$. Hence, j is the potential maximizer in a group with at least two potential maximizers, and so it is still of interest according to Algorithm 1.

- Conversely, suppose that j is of interest according to Algorithm 1, i.e., $\{j, j'\} \subseteq M_t^{(G)}$ for some G and $j' \neq j$. Since each group size is $n-k+1$, both j and j' must be potentially top-k, and neither of the two can be certifiably i-th best for $1 \leq i \leq k$. Therefore, both (i) and (ii) above hold.

2) LUCB: Recall that in each round of Algorithm 2, we select two arms $h_i := h_i^{G_i'}$ and $l_i := l_i^{G_i'}$, terminate if the former's LCB exceeds the latter's UCB, and otherwise pull both arms and continue. In the special case of top-k ranking, we claim that these steps can be reformulated as follows:

- In round i, find the arms h_i and l_i that maximize $\mathrm{UCB}_t(l_i) - \mathrm{LCB}(h_i)$ subject to the following constraints:
  1) h_i has one of the k highest empirical means (i.e., values of $\hat{\mu}_{j,T_j(t)}$) among all arms;
  2) The empirical mean of h_i is at least as high as that of l_i.
- If $\mathrm{UCB}_t(l_i) \leq \mathrm{LCB}(h_i)$ then terminate; otherwise pull both h_i and l_i and proceed to the next round.

Once again, this procedure can be efficiently implemented by sorting the relevant empirical means, UCB scores, and LCB scores, without the need to search over a combinatorially large number of groups.

To see the equivalence of the above steps to those in Algorithm 2, we note the following:

- Consider any pair $(h_i^{(G)}, l_i^{(G)})$ constructed in lines 4 and 5 of Algorithm 2. The fact that the empirical mean of h_i is at least as high as that of l_i is trivial, and since the group size is $n-k+1$, we also see that h_i has one of the k highest empirical means. Therefore, this pair is feasible in the maximization problem described above.
- Conversely, suppose that h_i and l_i are optimal (and therefore feasible) in the maximization problem described above. Consider a group G containing h_t, l_t, and an arbitrary set of $n-k-1$ other arms whose empirical mean is below that of h_t. By the optimality assumption, l_i must equal $\arg\max_{j \in G \setminus \{h_i^{(G)}\}} \mathrm{UCB}_t(j)$. and we deduce that this pair is indeed considered in lines 4 and 5 of Algorithm 2.

Since both variants of of the algorithm seek to maximize $\mathrm{UCB}_t(l_i) - \mathrm{LCB}(h_i)$ and terminate when this difference is non-positive, we deduce that the two are equivalent.

ACKNOWLEDGMENTS

This work was partially supported by the Swiss National Science Foundation (SNSF) under grant number 407540_167319, by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no725594 - time-data), and by an NUS Early Career Research Award.

REFERENCES

[1] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, (to appear). [Online]. Available: http://downloads.tor-lattimore.com/banditbook/book.pdf
[2] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges," *Statistical Science*, vol. 30, no. 2, 2015.
[3] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Int. Conf. World Wide Web*, 2010, pp. 661–670.
[4] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *ACM Symp. Theory Comp. (STOC)*, 2004, pp. 45–53.
[5] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *Int. Joint. Conf. Art. Intel. (IJCAI)*, vol. 15, 2015, pp. 974–980.
[6] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck, "Multi-bandit best arm identification," in *Conf. Neur. Inf. Proc. Sys. (NIPS)*, 2011, pp. 2222–2230.
[7] S. Bubeck and N. Cesa-Bianchi, *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*, ser. Found. Trend. Mach. Learn. Now Publishers, 2012.
[8] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. App. Math.*, vol. 6, no. 1, pp. 4 – 22, 1985.
[9] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Int. Conf. Comp. Learn. Theory*, 2002, pp. 255–270.
[10] J.-Y. Audibert and S. Bubeck, "Best arm identification in multi-armed bandits," in *Conf. Learning Theory (COLT)*, 2010.
[11] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," in *Int. Conf. Mach. Learn. (ICML)*, 2013.
[12] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, no. June, pp. 623–648, 2004.
[13] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Conf. Alg. Learn. Theory*, 2009.
[14] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits." in *Int. Conf. Mach. Learn. (ICML)*, 2012.
[15] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *J. Mach. Learn. Res. (JMLR)*, vol. 17, no. 1, pp. 1–42, 2016.
[16] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *Conf. Inf. Sci. Sys. (CISS)*, 2014.
[17] L. Chen, J. Li, and M. Qiao, "Nearly instance optimal sample complexity bounds for top-k arm selection," 2017, https://arxiv.org/abs/1702.03605.
[18] H. Jiang, J. Li, and M. Qiao, "Practical algorithms for best-k identification in multi-armed bandits," 2017, http://arxiv.org/abs/1705.06894.
[19] R. Heckel, N. B. Shah, K. Ramchandran, and M. J. Wainwright, "Active ranking from pairwise comparisons and when parametric assumptions don't help," *http://arxiv.org/abs/1606.08842*, 2016.
[20] S. Mohajer, C. Suh, and A. Elmahdy, "Active learning for top-k rank aggregation from noisy comparisons," in *Int. Conf. Mach. Learn. (ICML)*, 2017.
[21] S. Katariya, L. Jain, N. Sengupta, J. Evans, and R. Nowak, "Adaptive sampling for coarse ranking," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2018, pp. 1839–1848.
[22] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'UCB: An optimal exploration algorithm for multi-armed bandits," in *Conf. Learn. Theory (COLT)*, 2014.