Overlapping Multi-Bandit Best Arm Identification

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

Ilija Bogunovic
LIONS, EPFL
ilija.bogunovic@epfl.ch

Volkan Cevher
LIONS, EPFL
volkan.cevher@epfl.ch

***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.**

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.

A. Related Work

The literature on theory and algorithms for MAB problems is extensive; see [1], [7] for recent overviews. 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].

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.

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 establish a near-matching algorithm-independent 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 the full version [19], we show that the top-k ranking problem with regular bandit feedback (e.g., see [20]) is also a special case of our framework.

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 (after subtracting the mean) with parameter $\sigma \le \frac{1}{2}$; 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 $P_{\rm e} = \mathbb{P}\big[\bigcup_{G \in \mathcal{G}} \{\widehat{j}(G) \neq j^*(G)\}\big]$. We are interested in algorithms that achieve $P_{\rm e} \le \delta$ 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:

$$\Delta_j = \min\bigg\{ \min_{G \,:\, j \in G, j = j^*(G)} \big(\mu_j - \mu_{j_{\rm sec}(G)}\big), \\ \min_{G \,:\, j \in G, j \neq j^*(G)} \big(\mu_{j^*(G)} - \mu_j\big) \bigg\} \ge 0, \tag{2}$$

where $j_{\rm sec}(G)$ is the second-best arm in G, and the minimum of an empty set is infinity. 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 we assume this throughout the paper.

A. Auxiliary Results

Here we review some useful auxiliary results from the MAB literature that we will use in our analysis. We start with a useful concentration bound based on the law of iterated logarithm [21].

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 $\sigma \le \frac{1}{2}$. 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*

$$\left| \frac{1}{t}\sum_{s=1}^{t} Z_s - \mu \right| \le U(t,\delta), \quad \forall t \ge 1, \tag{3}$$

where

$$U(t,\delta) = (1+\sqrt{\epsilon})\sqrt{\frac{1+\epsilon}{2t}\log\frac{\log(1+\epsilon)t}{\delta}}. \tag{4}$$

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

$$\mathrm{UCB}_t(j) = \hat{\mu}_{j,T_j(t)} + U(T_j(t), \delta/n) \tag{5}$$
$$\mathrm{LCB}_t(j) = \hat{\mu}_{j,T_j(t)} - U(T_j(t), \delta/n), \tag{6}$$

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

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*

$$\mathrm{LCB}_t(j) \le \mu_j \le \mathrm{UCB}_t(j), \quad \forall j \in \{1, \dots, n\}, t \ge 1. \tag{7}$$

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, we use $n \cdot \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta/n}{\log(1+\epsilon)}\big)^{1+\epsilon} \le \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$. □

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.

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

$$\min\Big\{k \,:\, U(k, \delta/n) \le \frac{\Delta}{4}\Big\} \le \frac{2\gamma}{\Delta^2}\log\frac{2\log\big(\gamma(1+\epsilon)\Delta^{-2}\big)}{\delta/n}, \tag{8}$$

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 on T. 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*

$$\sum_{j=1}^{n} \mathbb{E}_\nu[N_j(\sigma)] D(\nu_j \| \nu'_j) \ge d(\mathbb{P}_\nu[\mathcal{A}], \mathbb{P}_{\nu'}[\mathcal{A}]), \tag{9}$$

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

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 $P_{\rm e} \le \delta$. 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

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

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 first main result is given as follows.

Theorem 1. (Lower bound) *Under Assumption 1, suppose that a given algorithm* $\texttt{Alg}^*$ *achieves* $P_{\rm e} \le \delta$ *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 arbitrarily small* $\alpha > 0$*:*

- *If the outer minimum in* (2) *is achieved by the first term (i.e., a group where* j *is best) then* $\mu_j' = \mu_j - (1+\alpha)\Delta_j$*;*
- *Otherwise,* $\mu_j' = \mu_j + (1+\alpha)\Delta_j$*.*

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

$$T_{\rm lower}(\delta) = \sum_{j=1}^{n} \frac{\log \frac{1}{2.4\delta}}{D(\nu_j \| \nu_j')}. \tag{10}$$

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 appendix of the full version [19], we provide the required changes to circumvent this assumption; these changes use the final part of Assumption 1. Letting $\mathcal{A}$ in (9) 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

$$\mathbb{E}_\nu[N_j] \ge \frac{\log \frac{1}{2.4\delta}}{D(\nu_j \| \nu_j')}. \tag{11}$$

Indeed, this follows from the fact that $P_{\rm e} \le \delta$ on all identifiable instances, and since by construction the KL divergence for arms indexed by $j' \ne j$ is zero (i.e., the distributions are identical in the two instances). Since (11) 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 (10). □

Remark 1. The bound (10) takes the same form as our upper bounds (to be given in the subsequent sections) whenever $D(\nu_j \| \nu_j') \le c\Delta_j^2$ for some constant c, in which case

$$T_{\rm lower}(\delta) \ge \sum_{j=1}^{n} \frac{\log \frac{1}{2.4\delta}}{c\Delta_j^2}. \tag{12}$$

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 $D(\nu_j \| \nu_j') \le \frac{\Delta_j^2(1+\alpha)^2}{\eta(1-\eta)}$ [7, Eq. (2.8)].

Algorithm 1 Successive Elimination Algorithm

```
Require: Groups $\mathcal{G}$, constants $\delta, \epsilon > 0$
Initialize $i = 1$, $t = 0$, and $T_j(t) = 0$ $(\forall j)$
Set $M_0^{(G)} = G$ $(\forall G)$, $\widetilde{\mathcal{G}}_0 = \mathcal{G}$, and $\mathcal{A}_0 = \{1, \dots, n\}$
while $\mathcal{A}_{i-1} \ne \emptyset$ do
    Pull every arm in $\mathcal{A}_{i-1}$ once, incrementing $t$ after each pull and updating all $T_j(t)$
    Compute $M_i^{(G)}$, $\widetilde{\mathcal{G}}_i$ and $\mathcal{A}_i$ via (13)–(15)
    For all $G$ with $|M_i^{(G)}| = 1$, set $\widehat{j}(G)$ to be the corresponding single arm.
    Increment the epoch index $i$
end while
return $(\widehat{j}(G_1), \dots, \widehat{j}(G_m))$
```

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 not below the top LCB:
$$M_i^{(G)} = \Big\{ j \in G \,:\, {\rm UCB}_{t_i}(j) \ge \max_{j' \in G} {\rm LCB}_{t_i}(j') \Big\} \tag{13}$$
under the definitions (5)–(6), 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:
$$\widetilde{\mathcal{G}}_i = \big\{ G \in \mathcal{G} \,:\, |M_i^{(G)}| \ge 2 \big\}, \tag{14}$$
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:
$$\mathcal{A}_i = \big\{ j \,:\, \exists G \in \widetilde{\mathcal{G}}_i \text{ with } j \in G \big\}. \tag{15}$$

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* $T_{\rm elim}(\delta, \epsilon)$ *arm pulls, where*

$$T_{\rm elim}(\delta, \epsilon) = \sum_{j=1}^{n} \frac{2\gamma}{\Delta_j^2} \log \frac{2\log\big(\gamma(1+\epsilon)\Delta_j^{-2}\big)}{\delta/n}, \tag{16}$$

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 $\widehat{j}(G) = j^*(G)$. 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

$$\begin{aligned} \mathrm{UCB}_{t_i}(j^*(G)) &\ge \mu_{j^*(G)} && (17) \\ &= \max_{j' \in G} \mu_{j'} && (18) \\ &\ge \max_{j' \in G} \mathrm{LCB}_{t_i}(j'), && (19) \end{aligned}$$

where both (17) and (19) use the validity of the confidence bounds (Corollary 1). Therefore, $j^*(G)$ meets the condition (13), 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 (8). 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 (5)–(6), $|\mathrm{UCB}_{t_i}(j') - \mathrm{LCB}_{t_i}(j')| < \frac{\Delta_j}{2}$. Hence,

$$\begin{aligned} \mathrm{UCB}_{t_i}(j') &< \mathrm{LCB}_{t_i}(j') + \frac{\Delta_j}{2} && (20) \\ &\le \mu_{j'} + \frac{\Delta_j}{2} && (21) \\ &\le \mu_j - \frac{\Delta_j}{2} && (22) \\ &\le \mathrm{UCB}_{t_i}(j) - \frac{\Delta_j}{2} && (23) \\ &< \mathrm{LCB}_{t_i}(j), && (24) \end{aligned}$$

  where (20) and (24) use the above-mentioned gap between UCB and LCB, (21) and (23) use the validity of the confidence bounds, and (22) uses the definition of Δ_j and the fact that $j = j^*(G)$. We see from (24) 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.
- On the other hand, if $j \in G$ is not the top arm in G, and if $U_i < \frac{\Delta_j}{4}$, a similar argument (detailed in the full version [19]) gives $\mathrm{UCB}_{t_i}(j) < \mathrm{LCB}_{t_i}(j^*(G))$. This implies that j is removed from $M_i^{(G)}$.

Combining these, we conclude that arm j only ever continues being pulled if $U_i \ge \frac{\Delta_j}{4}$. Since i is precisely the number of arm pulls of all remaining arms after epoch i, applying Lemma 2 and summing (8) over $j = 1, \dots, n$ yields (16). □

Algorithm 2 LUCB-Type Algorithm

```
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 $w_i^{(G'_i)} \le 0$ then
        return $(h_i^{(G_1)}, \dots, h_i^{(G_m)})$
    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
```

We observe that (16) matches (12) 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 (e.g., we in fact have tightness up to constant factors when $\delta = O(n^{-\alpha})$ and $\min_j \Delta_j = \Omega(n^{-\beta})$ for some constants $\alpha, \beta > 0$).

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 $i \ge 1$.

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* $\delta \in \big(0, \frac{1}{e}\log(1+\epsilon)\big)$, *with probability at least* $1 - \frac{2+\epsilon}{\epsilon/2}\big(\frac{\delta}{\log(1+\epsilon)}\big)^{1+\epsilon}$, *the LUCB algorithm terminates with the correct output after at most* $T_{\mathrm{lucb}}(\delta, \epsilon)$ *arm pulls, where*

$$T_{\mathrm{lucb}}(\delta, \epsilon) = 2 \sum_{j=1}^{n} \frac{2\gamma}{\Delta_j^2} \log \frac{2\log\big(\gamma(1+\epsilon)\Delta_j^{-2}\big)}{\delta/n}, \qquad (25)$$

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

Observe that this result matches that of Theorem 2 up to a factor of 2, and is therefore similarly near-optimal with respect to the algorithm-independent lower bound of Theorem 1.

Proof outline. We first show that when the high probability event in Corollary 1 holds, the algorithm can only terminate with the correct output $(j^*(G_1),\dots,j^*(G_m))$. Suppose for the purpose of contradiction that the algorithm terminates during round i and returns $(h_i^{(G_1)},\dots,h_i^{(G_m)}) \neq (j^*(G_1),\dots,j^*(G_m))$. 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

$$\begin{aligned}
\mu_{h_i^{(G)}} &\geq \mathrm{LCB}_{t_i}(h_i^{(G)}) && (26)\\
&\geq \mathrm{UCB}_{t_i}(l_i^{(G)}) && (27)\\
&\geq \mathrm{UCB}_{t_i}(j^*(G)) && (28)\\
&\geq \mu_{j^*(G)}, && (29)
\end{aligned}$$

where (26) and (29) use the validity of the confidence bounds (Corollary 1), (27) uses the stopping condition, and (28) uses the definition of $l_i^{(G)}$. From (29), 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.

Bounding the number of pulls of each arm requires more effort, and the detailed are deferred to the full version [19] due to space constraints. The main steps are as follows:

(Step 1) 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, and 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:

$$\begin{aligned}
&j = j^*(G) \text{ and } \mathrm{LCB}_{t_i}(j) < c(G), \text{ or} && (30)\\
&j \neq j^*(G) \text{ and } \mathrm{UCB}_{t_i}(j) > c(G). && (31)
\end{aligned}$$

Conditioned on the event in Corollary 1, we show that

$$\begin{aligned}
&\mathrm{LCB}_{t_i}(h_i(G'_i))) < \mathrm{UCB}_{t_i}(l_i(G'_i))) \implies\\
&\{h_i(G'_i) \text{ is } G'_i\text{-}\textit{BAD}\} \text{ or } \{l_i(G'_i) \text{ is } G'_i\text{-}\textit{BAD}\} \qquad (32)
\end{aligned}$$

for all $i \geq 1$. That is, if the stopping condition is not satisfied then either $h_i(G'_i)$ or $l_i(G'_i)$ is G'_i-*BAD*.

(Step 2) 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. This is shown separately for the cases $j \notin \{j^*(G_1),\dots,j^*(G_m)\}$ and $j \in \{j^*(G_1),\dots,j^*(G_m)\}$.

(Step 3) The preceding steps can be combined to deduce that conditioned on the high probability event from Corollary 1 the total number of rounds does not exceed the following:

$$\begin{aligned}
&\sum_{i=1}^{\infty} \mathbf{1}\{h_i^{(G'_i)} \text{ is } G'_i\text{-}\textit{BAD} \text{ or } l_i^{(G'_i)} \text{ is } G'_i\text{-}\textit{BAD}\}\\
&\leq \sum_{j=1}^{n} (\tau_j - 1), \qquad (33)
\end{aligned}$$

from which Theorem 3 follows by noting that the total number of arm pulls is $n + (2 \times \text{number of rounds})$, and substituting the upper bound in (8) (with $\Delta = \Delta_j$) for τ_j.

□

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] J. Scarlett, I. Bogunovic, and V. Cevher, "Overlapping multi-bandit best arm identification (technical report)," 2019, https://infoscience.epfl.ch/record/265112/files/MultiBandit_FULL.pdf.

[20] 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.

[21] 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.