Best Arm Identification in Bandits with Limited Precision Sampling

Kota Srinivas Reddy*, P. N. Karthik†, Nikhil Karamchandani‡, and Jayakrishnan Nair‡
*Indian Institute of Technology Madras, † National University of Singapore, ‡ Indian Institute of Technology Bombay
Emails: ksvr1532@gmail.com, karthik@nus.edu.sg, nikhilk@ee.iitb.ac.in, jayakrishnan.nair@ee.iitb.ac.in

***Abstract*—We study best arm identification in a variant of the multi-armed bandit problem where the learner has limited precision in arm selection. The learner can only sample arms via certain exploration bundles, which we refer to as boxes. In particular, at each sampling epoch, the learner selects a box, which in turn causes an arm to get pulled as per a box-specific probability distribution. The pulled arm and its instantaneous reward are revealed to the learner, whose goal is to find the best arm by minimising the expected stopping time, subject to an upper bound on the error probability. We present an asymptotic lower bound on the expected stopping time, which holds as the error probability vanishes. We show that the optimal allocation suggested by the lower bound is, in general, non-unique and therefore challenging to track. We propose a modified tracking-based algorithm to handle non-unique optimal allocations, and demonstrate that it is asymptotically optimal. We also present non-asymptotic lower and upper bounds on the stopping time in the simpler setting when the arms accessible from one box do not overlap with those of others.**

For a longer version of this paper with proofs, see [1].

I. Introduction

In this paper, we study best arm identification in a multi-armed bandit setting with K arms, where the learner has limited precision in sampling arms. In particular, the learner cannot directly sample individual arms, but can instead sample only certain exploration bundles, which we refer to as *boxes*. Each box is associated with a probability distribution over the arms; upon selecting a box, an arm is pulled randomly according to its corresponding probability distribution. The learner sees the pulled arm and its instantaneous reward. The learner's goal is to find the *best arm*, defined as the arm with the largest mean reward, while minimising the expected stopping time subject to an upper bound on the error probability (i.e., *fixed-confidence* regime). For the *boxed-bandit best arm identification* problem described above, our objective is to design sound algorithms and benchmark their performance against information theoretic lower bounds, when the arm reward distributions and the arm selection probabilities of the various boxes are apriori unknown.

A. Motivation

The key feature of our model is that the learner does not have direct access to the arms. Instead, it must perform its exploration via certain intermediaries (boxes), which have their own preferences/biases/constraints over arm selection. To consider a contemporary example, suppose the goal of the learner is to identify the most contagious strain of a virus/pathogen in a large community (say a country), by ordering tests at different local testing facilities. Each testing facility in turn performs its tests by sampling individuals in its local vicinity/jurisdiction; the likelihood of encountering different strains being a function of the facility location.

Another interpretation of our model is that it captures *noise* in arm selection. For example, when the learner attempts to pull a certain arm, the pull is only executed successfully with, say probability $1-\eta$; with probability η, either no arm is pulled (i.e., an *erasure* occurs), or a random arm is pulled (this is the *trembling hand* model of [2]).

Finally, our model can also be interpreted as a *privacy preservation* exercise on the part of learner. By performing its exploration via the *non-adaptive* selection profiles of the boxes, the learner can obfuscate its own preferences from other observers. Naturally, this obfuscation comes at the expense of increased sampling complexity. Under this alternative interpretation, it may be reasonable to assume that the learner knows the arm selection probabilities of the boxes; our algorithms simplify naturally to this special case.

B. Analytical Challenges

Notice that in our problem setup, the learner only has *partial* control over the arms (via the boxes). This is unlike classical best arm identification problems [3]–[5] where the learner has full control over the arm to pull at each time instant. For instance, in the Successive Elimination algorithm of [6] or the LUCB algorithm of [7], the learner pulls one or more arms at each time instant and either eliminates the sub-optimal arms or resolves between the best and second-best arms on-the-fly to eventually arrive at the best arm. In our setup, because a given arm may be accessible via multiple boxes, and the arm selection probabilities of the boxes are not known beforehand, it is not clear at the outset which box must be selected more frequently to maximise the chances of pulling an arm. In fact, if each arm belongs to every box and the arms selection probabilities of the boxes are all identical, then every randomised box selection rule yields the same expected stopping time. Thus, elimination or LUCB-type algorithms do not apply verbatim to our setting.

We also note that the lower bounds appearing in [4], [5] admit a unique optimal solution (or *allocation*), and a key aspect of the best arm identification algorithms in these works is *tracking* or the convergence of the empirical arm selection frequencies to the optimal allocation. In contrast, we show that

978-1-6654-7554-9/23/$31.00 ©2023 IEEE

the optimal allocation in our setup is in general non-unique. In this case, the empirical frequencies may alternate between two or more optimal allocations and not converge to any of the optimal allocations in the long run. This underscores the need to improvise the existing tracking-based algorithms of [4], [5] to handle non-unique optimal allocations.

C. Contributions

We derive an asymptotic lower bound on the growth rate of the expected stopping time, where the asymptotics is as the error probability vanishes. We show that this growth rate is captured by a sup-inf optimisation problem whose optimal (sup-attaining) solution (or *allocation*) is potentially non-unique. Inspired from the analytical techniques of Jedra et al. [8], we propose a tracking-based algorithm that is improvised to handle non-unique allocations at every time step, and demonstrate that our algorithm achieves the lower bound asymptotically. In our achievability analysis, we track the long-term behaviour of the empirical average of all the past allocations, and show that the mean allocation eventually approaches the correct set of optimal allocations. Our achievability analysis can potentially be applied to more general problem settings where non-unique allocations arise naturally as in our work, or where proving the uniqueness of the optimal allocation is hard; see, for instance, the remarks in [2], [9].

Finally, in the special case when the arms are *partitioned* among the boxes, i.e., when set of arms accessible from one box does not overlap with that of the others, we present non-asymptotic guarantees for a variant of the successive elimination algorithm. We show that the expected stopping time of this algorithm satisfies an upper bound that is tight in the unknown instance parameters.

D. Related Works

Our setup closely resembles that in [10], where the arms are grouped into subsets as in our work, with potential overlap between the sets. However, the key difference is that in [10], the learner has full control over the arms (unlike partial control of arms in our work). Also, in [10], the goal is to find the best arm within each subset, whereas in our work, the goal is to find the overall best arm. The paper [11] considers a similar setup as ours for a problem of community mode estimation, with the key difference that the analysis and results in [11] are for the *fixed-budget* regime, whereas those of our work are for the *fixed-confidence* regime; see [4] for a comparison of these regimes. Our setup specialises to those in [3], [4], [6], [7] when the number of boxes equals the number of arms, and each box contains one arm. Our setup also specialises to the *trembling hand model* of [2]; in this model, when the learner attempts to pull arm k, it actually gets pulled with probability $1-\eta$, whereas a random arm, chosen uniformly, gets pulled with probability η.

II. Formulation and Preliminaries

We consider a K-armed bandit, where the arms are labelled $1, 2, \ldots, K$. Arm $k \in [K]$[1] is associated with a reward distribution $\nu_k \in \mathcal{G}$, where $\mathcal{G}$ is a known class of arm distributions. Let μ_k denote the mean reward of arm k. The goal of the learner is to identify, via sequential sampling, the optimal arm, which is defined to be the arm having the largest mean reward. However, unlike in the classical MAB setting, the learner cannot sample (a.k.a., pull) individual arms directly. Instead, at each epoch, the learner selects a *box* (from a finite collection of M boxes), which results in an arm being pulled randomly according to a box-dependent probability distribution. Formally, selecting box $m \in [M]$ results in arm $k \in \mathcal{A}_m \subseteq [K]$ being pulled with probability $q_{m,k}$. Here, $\mathcal{A}_m$ denotes the set of arms that are 'accessible' using box m (i.e., $q_{m,k} > 0$ for $k \in \mathcal{A}_m$ and $\sum_{k\in\mathcal{A}_m} q_{m,k} = 1$). Importantly, $\boldsymbol{q} \coloneqq \{q_{m,k} : k \in \mathcal{A}_m, m \in [M]\}$ is apriori unknown to the learner. Note that $\boldsymbol{q}$ describes the imprecision in the learner's ability to pull specific arms. Indeed, capturing the impact of this sampling imprecision on the complexity of best arm identification is the main focus of this work.

The tuple $C = (\boldsymbol{q}, \boldsymbol{\nu})$ completely specifies a *problem instance*, where $\boldsymbol{\nu} = (\nu_k : k \in [K])$ is the vector of arm distributions. The optimal arm corresponding to this problem instance is denoted by $a^\star(C) = a^\star(\boldsymbol{\mu}) = \arg\max_{k\in[K]} \mu_k$, where $\boldsymbol{\mu} = (\mu_k : k \in [K])$. The best arm is assumed to be uniquely defined for every problem instance. We write $\textsc{Alt}(\boldsymbol{\mu})$ to denote the set of instances *alternative* to $\boldsymbol{\mu}$, i.e., those instances whose best arm differs from $a^\star(\boldsymbol{\mu})$. When there is no ambiguity, we write $C = (\boldsymbol{q}, \boldsymbol{\mu})$ in place of $C = (\boldsymbol{q}, \boldsymbol{\nu})$.

For $t \in \{1, 2, \ldots\}$, let B_t denote the box selected by the learner at time t. Upon selecting box $B_t = m$, arm $A_t = k$ is pulled with probability $q_{m,k}$. The learner observes A_t (i.e., it knows which arm was pulled) and the reward X_t from arm A_t. Let $(B_{1:t}, A_{1:t}, X_{1:t}) \coloneqq (B_1, A_1, X_1, \ldots, B_t, A_t, X_t)$ denote the history of box selections, arm pulls, and observations seen up to time t. Given $\delta \in (0, 1)$, the goal of the learner is to find the best arm with the least expected number of box selections (a.k.a. expected stopping time), while keeping the stoppage error probability below δ.

Let $\pi = \{\pi_t\}_{t=1}^{\infty}$ denote any generic best arm identification *policy* (or algorithm), where for every $t \geq 1$, π_t maps the history $(B_{1:t}, A_{1:t}, X_{1:t})$ to one of the following actions:

- Select box B_{t+1} according to a deterministic or randomised rule.
- Stop and declare the estimated best arm.

Let τ_π denote the (random) stopping time under π, and let $\hat{a}$ be the best arm estimate at stoppage. For each $\delta \in (0, 1)$, our interest is in the class of δ-*probably correct* (δ-PC) algorithms defined by $\Pi(\delta) \coloneqq \{\pi : P(\hat{a} \neq a^\star(C)) \leq \delta \quad \forall C\}$. In this paper, we design δ-PC algorithms and benchmark their stopping times against information theoretic lower bounds.

In Section III, we design and analyse a track-and-stop style algorithm taking the class of arm distributions $\mathcal{G}$ to be the family of Gaussian distributions with a known variance. This algorithm is shown to be δ-PC, and its expected stopping time is shown to be asymptotically optimal as $\delta \downarrow 0$. Next, in Section IV, we consider the special case of our model where the sets $\{\mathcal{A}_m : m \in [M]\}$ are disjoint (i.e., the arms are

[1]Let $[n] \coloneqq \{1, \ldots, n\}$ for any integer $n \geq 1$.

partitioned across boxes). For this case, taking $\mathcal{G}$ to be the family of 1-sub-Gaussian distributions, we design and analyse a successive-elimination style algorithm which admits non-asymptotic (in δ) stopping time bounds.

III. Track & Stop Based Algorithm

We first study the general setting when each arm may be associated with multiple boxes. For simplicity in presentation, we assume that the observations from arm k are Gaussian distributed with mean μ_k and variance 1. Without loss of generality, we present our results for the extreme setting when each arm is associated with *every* box. Let the underlying instance be defined by $\boldsymbol{q}_0 = \{q^0_{m,k} : m \in [M], k \in [K]\}$ and $\boldsymbol{\mu}_0 = \{\mu^0_k : k \in [K]\}$. We first present an asymptotic (as $\delta \downarrow 0$) lower bound on the growth rate of the expected stopping time. Following this, we highlight the central challenge in the analysis of the non-partition setting: *non-uniqueness* of the optimal solution to the optimization problem that characterizes the lower bound. We then present a new track-and-stop based algorithm inspired from [8] and demonstrate its asymptotic optimality despite the above challenge.

A. Converse: Asymptotic Lower Bound

The first main result of this section, a lower bound on the limiting growth rate of the expected stopping time for δ-PC algorithms in the limit as $\delta \downarrow 0$, is presented below.

Theorem 1: Let $\boldsymbol{q}_0 = \{q^0_{m,k}\}_{m,k}$, $\boldsymbol{\mu}_0 = \{\mu^0_k\}_{k=1}^K$. Then,

$$\liminf_{\delta\downarrow 0} \inf_{\pi\in\Pi(\delta)} \frac{\mathbb{E}[\tau_\pi]}{\log(1/\delta)} \geq \frac{1}{T^*(\boldsymbol{q}_0,\boldsymbol{\mu}_0)}, \tag{1}$$

where $T^*(\boldsymbol{q}_0,\boldsymbol{\mu}_0)$ in (1) is given by

$$T^*(\boldsymbol{q}_0,\boldsymbol{\mu}_0) = \sup_{w\in\Sigma_M} \inf_{\boldsymbol{\lambda}\in \mathrm{ALT}(\boldsymbol{\mu}_0)} \sum_{m=1}^M \sum_{k=1}^K w_m\, q^0_{m,k}\, \frac{(\mu^0_k-\lambda_k)^2}{2}, \tag{2}$$

where Σ_M is the simplex of all probability distributions (or *allocations*) $w = (w_1,\ldots,w_M)$ on the boxes.

The proof of Theorem 1 is quite standard and omitted for brevity. It employs a change-of-measure argument for bandits [12], the transportation lemma of [4], and Wald's identity.

B. Non-Uniqueness of the Optimal Allocation

Consider the following simple example with $M = 2$ boxes and $K = 4$ arms. Suppose that $\boldsymbol{\mu}_0 = \{0.5, 0.4, 0.3, 0.3\}$. Notice that arm 1 is the best arm. Let $\boldsymbol{q}_0$ be specified by the following matrix:

$$\boldsymbol{q}_0 = \begin{pmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix}.$$

The first row of the above matrix represents the arm selection probabilities of box 1, and the second row that of box 2. For this example, it is easy to show that every $w = (w_1, w_2) \in \Sigma_2$ attains the supremum in (2) and hence is an optimal allocation.

The above example shows that the optimal allocation can potentially be non-unique. This is in contrast to the prior works [4], [5] where the optimal allocation is unique. Let $\mathcal{W}^\star(\boldsymbol{q}_0,\boldsymbol{\mu}_0)$ denote the set of all allocations that attain the supremum in (2). More generally, let $\mathcal{W}^\star(\boldsymbol{q},\boldsymbol{\mu})$ denote the set of optimal allocations corresponding to the instance $(\boldsymbol{q},\boldsymbol{\mu})$.

Lemma 1: The mapping $(\boldsymbol{q},\boldsymbol{\mu}) \mapsto \mathcal{W}^\star(\boldsymbol{q},\boldsymbol{\mu})$ is upper-hemicontinuous and compact-valued. Furthermore, $\mathcal{W}^\star(\boldsymbol{q},\boldsymbol{\mu})$ is convex for all $(\boldsymbol{q},\boldsymbol{\mu})$.

In particular, Lemma 1 implies that $\mathcal{W}^\star(\boldsymbol{q}_0,\boldsymbol{\mu}_0)$ is convex; this, we shall see, will play an important role in the design of an asymptotically optimal algorithm, which forms the content of the next section.

C. Achievability: Handling Non-Unique Allocations

A key feature of the best arm identification algorithms and the contingent achievability analyses in the prior works [4], [5] is *tracking*, or the almost sure convergence of the empirical frequencies of arm pulls to the optimal allocation. When the optimal allocation is non-unique as in our work, the empirical frequencies may alternate among two or more optimal allocations in the long-run and not converge to any one of the optimal allocations, in which case it is difficult to establish tracking. This emphasises the need to improvise the existing tracking-based algorithms to handle potentially non-unique optimal allocations and develop a framework to demonstrate tracking-like behaviour. In this section, we present a technique that achieves this.

Let $N(t,m,k)$ be the total number of times box m is selected and arm k is pulled up to time t. Define $N(t,m) = \sum_k N(t,m,k)$ be the number of times box m is selected up to time t, and $N_k(t) = \sum_m N(t,m,k)$ be the number of times arm k is pulled up to time t. For all m,k, let

$$\hat{q}_{m,k}(t) = \frac{N(t,m,k)}{N(t,m)}, \quad \hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^t \mathbf{1}_{\{A_s=k\}}\, X_s, \tag{3}$$

be the empirical estimates of the unknown parameters at time t. The key result (inspired from [8]) that enables us to prove achievability while working with a set of optimal allocations, is stated below.

Lemma 2: Let $f(t) = \frac{\sqrt{t}}{\sqrt{M}}$. Let $\{w(t)\}_{t=1}^\infty \subset \Sigma_M$ be any sequence such that $w(t+1) \in \mathcal{W}^\star(\hat{\boldsymbol{q}}(t),\hat{\boldsymbol{\mu}}(t))$ for all t. Let $i_0 = 0$ and

$$i_{t+1} = (i_t \bmod M) + \mathbf{1}_{\left\{\min_{m\in[M]} N(t,m) < f(t)\right\}}, \quad t \geq 0.$$

Then, under the *modified D-tracking rule* given by

$$B_{t+1} = \begin{cases} i_t, & \min_{m\in[M]} N(t,m) < f(t), \\ b_t, & \text{otherwise}, \end{cases} \tag{4}$$

where $b_t = \arg\min_{m\in \mathrm{supp}(\sum_{s=1}^t w(s))} N(t,m) - \sum_{s=1}^t w_m(s)$, we have

$$\lim_{t\to\infty} d_\infty((N(t,m)/t)_{m\in[M]},\ \mathcal{W}^\star(\boldsymbol{q}_0,\boldsymbol{\mu}_0)) = 0 \quad \text{a.s..} \tag{5}$$

Notice that the sampling rule in (4) selects each of the boxes once at the beginning. At time t, if any one of the boxes is under-sampled, i.e., $\min_m N(t,m) < f(t)$, the boxes are sampled forcefully in a round-robin fashion until

Algorithm 1 Boxed-Bandit Modified Track-and-Stop

Input: $\delta \in (0,1)$, $\rho > 0$, $K \in \mathbb{N}$, and $M \in \mathbb{N}$
Output: $\hat{a} \in [K]$ – best arm
Initialisation: $t = 0$, $\hat{\mu}_a(t) = 0 \, \forall a \in [K]$, $Z(0) = 0$.
Compute $\hat{a}(t) = \arg\max_a \hat{\mu}_a(t)$.
if $Z(t) \geq \zeta(t, \delta, \rho)$ and $\min_k N_k(t) > 0$ **then**
  Stop box selections.
  return $\hat{a} = \arg\max_{a \in [K]} \hat{\mu}_a(t)$
else
  Select box B_{t+1} as per *modified D-tracking rule* (4).
  Update $\hat{\boldsymbol{q}}(t)$, $\hat{\boldsymbol{\mu}}(t)$, and $Z(t)$. Go to step 1.
end if

$N(t,m) = \Omega(\sqrt{t})$ for all m. Else, a box is sampled based the allocations $\{w(s) : 1 \leq s \leq t\}$. Here, $w(s)$ is an *arbitrary* element of $\mathcal{W}^\star(\hat{\boldsymbol{q}}(s-1), \hat{\boldsymbol{\mu}}(s-1))$, the set of optimal allocations corresponding to the *estimated* instance parameters at time $s-1$. This is the principle of *certainty equivalence.* In the proof, we show that the empirical average of all the allocations up to time t, $\bar{w}(t) = 1/t \sum_{s=1}^{t} w(s)$, approaches $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ as $t \to \infty$. Thanks to the forced exploration of the boxes, $(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t))$ approaches $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ a.s. as $t \to \infty$. This, together with the upper-hemicontinuity property of $\mathcal{W}^\star$ from Lemma 1, implies that $\mathcal{W}^\star(\boldsymbol{q}(t), \boldsymbol{\mu}(t))$ is "close" to $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ for large t, a.s., from which we arrive at (5).

Let $Z_{a,b}(t)$ denote the *generalised likelihood ratio test statistic* between arms $a, b \in [K]$ up to time t, defined as

$$Z_{a,b}(t) \coloneqq \log \frac{\sup_{(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \geq \mu_b} P(B_{1:t}, A_{1:t}, X_{1:t})}{\sup_{(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \leq \mu_b} P(B_{1:t}, A_{1:t}, X_{1:t})}. \quad (6)$$

The next result provides an explicit expression for $Z_{a,b}(t)$.

Lemma 3: Fix $a, b \in [K]$ and a policy π. Fix t such that $\min_{k \in [K]} N_k(t) > 0$ a.s.. If $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$, then

$$Z^{\pi}_{a,b}(t) = N_a(t) \frac{\left(\hat{\mu}_a(t) - \hat{\mu}_{a,b}(t)\right)^2}{2} + N_b(t) \frac{\left(\hat{\mu}_b(t) - \hat{\mu}_{a,b}(t)\right)^2}{2}, \quad (7)$$

where $\hat{\mu}_{a,b}(t)$ is defined as

$$\hat{\mu}_{a,b}(t) \coloneqq \frac{N_a(t)}{N_a(t) + N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t) + N_b(t)} \hat{\mu}_b(t). \quad (8)$$

If $\hat{\mu}_a(t) \leq \hat{\mu}_b(t)$, then $Z^{\pi}_{a,b}(t) = -Z^{\pi}_{b,a}(t)$. Thus, $Z^{\pi}_{a,b}(t) \geq 0$ if and only if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$.

Let $Z(t) = \max_a \min_{b \neq a} Z_{a,b}(t)$. For fixed $\delta \in (0,1)$ and $\rho > 0$, let $\zeta(t, \delta, \rho) \coloneqq \log(C\, t^{1+\rho}/\delta)$, where C is a constant that satisfies $\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} \frac{(\log^2(C\, t^{1+\rho}) \, \log t)^K}{t^{1+\rho}} \leq C$.

Algorithm for best arm identification: We propose a variant of the track-and-stop algorithm of Garivier et al. [5], called *Boxed-Bandit Modified Track-and-Stop* or BBMTS, that is improvised to work with a set of allocations at each time step. Our algorithm takes as input two parameters: $\delta \in (0,1)$ and $\rho > 0$. At each time instant t, the algorithm maintains an estimate of the best arm $\hat{a}(t) \in \arg\max_a \hat{\mu}_a(t)$, with ties resolved uniformly randomly. Lemma 3 shows that $\hat{a}(t) \in \arg\max_a \min_{b \neq a} Z_{a,b}(t)$. Then, the algorithm checks if $\min_{b \neq \hat{a}(t)} Z_{\hat{a}(t),b}(t) \geq \zeta(t, \delta, \rho)$. If this holds, the algorithm is sufficiently confident that $\hat{a}(t)$ is the best arm; it stops and outputs $\hat{a}(t)$ as the best arm. Else, the algorithm samples box B_{t+1} according to the modified D-tracking rule (4) as explained earlier. See Algorithm 1 for a pseudo-code.

Performance analysis of BBMTS: Let $\pi_{\text{BBMTS}}(\delta, \rho)$ symbolically represent the BBMTS algorithm with parameters δ, ρ. The below result characterises its performance.

Theorem 2: The BBMTS algorithm meets the following performance criteria.

1) $\pi_{\text{BBMTS}}(\delta, \rho) \in \Pi(\delta)$ for each $\delta \in (0,1)$ and $\rho > 0$.
2) For each $\rho > 0$, the stopping time of $\pi_{\text{BBMTS}}(\delta, \rho)$ satisfies
$$\limsup_{\delta \downarrow 0} \frac{\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}}{\log(1/\delta)} \leq \frac{1+\rho}{T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)} \quad \text{a.s..} \quad (9)$$
3) For each $\rho > 0$, the quantity $\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}]$ satisfies
$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}]}{\log(1/\delta)} \leq \frac{1+\rho}{T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)}. \quad (10)$$

Notice that ρ serves as a tuneable parameter that may be set to make the upper bound in (10) as close to (1) as desired. Clearly, the right-hand side of (10) matches with the lower bound in (1) as $\rho \downarrow 0$. Hence, Theorems 1 & 2 together imply that $1/T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ is the optimal asymptotic growth rate of the expected stopping corresponding to the instance $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$.

IV. Partition Setting

In this section, we analyze the simpler setting when the K arms are *partitioned* among the M boxes and the learner knows $\{\mathcal{A}_m : m \in [M]\}$ but not the arm selection probabilities of the boxes. For this setting, we present an algorithm based on successive elimination, and provide a non-asymptotic, high-probability upper bound on its stopping time. We show that the upper bound is tight in the instance-specific parameters.

In the partition setting, we find it convenient to index an arm by the box it is associated with; the kth arm in box m is denoted by $A_{m,k}$ and its mean reward is denoted by $\mu_{m,k}$. All arms are assumed to yield 1-sub-Gaussian rewards. Let $\boldsymbol{q}_0 = \{q^0_{m,k} : k \in \mathcal{A}_m, \, m \in [M]\}$ and $\boldsymbol{\mu}_0 = \{\mu^0_{m,k} : k \in \mathcal{A}_m, \, m \in [M]\}$ define the underlying instance. Without loss of generality, let $A_{1,1}$ be the best arm in this instance. We let $\Delta_{m,k} \coloneqq \mu^0_{1,1} - \mu^0_{m,k}$ for all $(m,k) \neq (1,1)$, and $\Delta_{1,1} \coloneqq \min_{(m,k) \neq (1,1)} \Delta_{m,k}$ denote the arm sub-optimality gaps.

A. Non-Asymptotic Analysis: Achievability

We now propose an algorithm based on successive elimination (called the *Boxed-Bandit Successive Elimination Algorithm* or BBSEA) and analyze its performance. Before presenting the algorithm, we introduce some notations. Our algorithm proceeds in rounds; we use n to denote the round number and t to denote the running time. Let $t_{m,k}(n)$ denote the number of times arm $A_{m,k}$ is pulled up to round n, and let $\hat{\mu}_{m,k}(n)$ denote the empirical mean of arm $A_{m,k}$ after $t_{m,k}(n)$ pulls. In any given round n, let S denote the set of *active arms* (candidate best arms), S_m the set of *active*

Algorithm 2 Boxed-Bandit Successive Elimination Algorithm

Input: K, M, $\delta > 0$, $\mathcal{A}_m$ for $m \in [M]$
Output: $\hat{a}_{\text{BBSEA}} \in [K]$ (best arm).
Initialization: $B = [M]$, $n = 0$, $t = 0$,
$S_m = \mathcal{A}_m \ \forall m$, $S = \bigcup_m S_m$, $\hat{\mu}_{m,k}(0) = 0 \ \forall m, k$.
while $|S| > 1$ **do**
  $n \leftarrow n + 1$
  For each $m \in B$, select box m until every active arm $A_{m,k}$ in box m is pulled at least n times. For every box selection, increment t by 1.
  Update $t_{m,k}(n)$, $\hat{\mu}_{m,k}(n)$, $\text{UCB}_{m,k}(n)$ and $\text{LCB}_{m,k}(n)$ for all the active arms.
  if $\exists A_{m',k'} \in S$ such that $\text{UCB}_{m,k}(n) < \text{LCB}_{m',k'}(n)$ **then**
    $S_m \leftarrow S_m \backslash A_{m,k}$, $S \leftarrow \bigcup_{m \in [M]} S_m$,
    $B \leftarrow \{m : S_m \neq \emptyset\}$.
  end if
  if $|S| = 1$ **then**
    $\hat{a}_{\text{BBSEA}} \leftarrow a \in S$, $S \leftarrow \emptyset$, $B \leftarrow \emptyset$.
  end if
end while
return $\hat{a}_{\text{BBSEA}}$.

arms associated with box m, and B the set of *active boxes* (boxes that are associated with at least one active arm). For a fixed $\delta \in (0,1)$, let $\alpha_\delta(x) := \sqrt{\frac{2\log(8Kx^2/\delta)}{x}}$. Let $\text{UCB}_{m,k}(n)$ and $\text{LCB}_{m,k}(n)$ denote respectively the upper and lower confidence bounds on $\hat{\mu}_{m,k}(n)$ with confidence interval of length $\alpha_\delta(t_{m,k}(n))$, i.e.,

$$\text{UCB}_{m,k}(n) = \hat{\mu}_{m,k}(n) + \alpha_\delta(t_{m,k}(n)), \tag{11}$$
$$\text{LCB}_{m,k}(n) = \hat{\mu}_{m,k}(n) - \alpha_\delta(t_{m,k}(n)). \tag{12}$$

Let $\hat{a}_{\text{BBSEA}}$ denote the best arm output by BBSEA. We set the initial values $t = 0$, $n = 0$, $S = [K]$, $B = [M]$, and $S_m = \mathcal{A}_m$ for all m. In each round n, the following sequence of actions is executed by BBSEA: (i) Each box $m \in B$ is selected until *every active arm* associated with box m has been pulled at least n times; for every box selection, time t is incremented by 1. (ii) With every box selection and arm pull, the values of $t_{m,k}(n)$, $\hat{\mu}_{m,k}(n)$, $\text{UCB}_{m,k}(n)$, and $\text{LCB}_{m,k}(n)$ are updated for all the active arms. (iii) Arm $A_{m,k}$ is eliminated from S in round n if $\text{UCB}_{m,k}(n) \leq \max_{m',k'} \text{LCB}_{m',k'}(n)$. The above sequence of actions repeats until only one arm remains in S, at which point the algorithm stops and outputs the single arm in S as the best arm. Because t is incremented with every box selection, the stopping time is equal to the total number of box selections. See Algorithm 2 for the pseudo-code of BBSEA.

Performance analysis of BBSEA: Before we present the results on the performance of BBSEA, we introduce a few useful notations. For each (m,k) pair and $\delta \in (0,1)$, let $\alpha_{m,k} = 1 + \frac{102}{\Delta_{m,k}^2} \log\left(\frac{64\sqrt{\frac{8K}{\delta}}}{\Delta_{m,k}^2}\right)$, and let $T_{m,k}$ denote the round number in which arm $A_{m,k}$ is eliminated from S (the set of active arms). Furthermore, let $\beta_{m,k} = \frac{1}{q_{m,k}^0}\left[\alpha_{m,k} + 2\log\frac{2K}{\delta} + 2\sqrt{\log\frac{2K}{\delta}\left(\log\frac{2K}{\delta} + \alpha_{m,k}\right)}\right]$, and let $\beta_m = \max_{k \in \mathcal{A}_m} \beta_{m,k}$.

With the above notations in place, we are now ready to state the main result of this section on the performance of BBSEA.

Theorem 3: Fix $\delta \in (0,1)$, The following hold with probability greater than $1 - \delta$.

1) BBSEA outputs the best arm correctly.
2) The stopping time of BBSEA is $\leq \sum_{m=1}^{M} \beta_m$.

Furthermore, for any $\pi \in \Pi(\delta)$,

$$\mathbb{E}[\tau_\pi] \geq \log\left(\frac{1}{2.4\,\delta}\right) \cdot \sum_{m=1}^{M} \max_{k \in \mathcal{A}_m} \frac{1}{q_{m,k}^0\,\Delta_{m,k}^2}. \tag{13}$$

Notice that

$$\beta_{m,k} = O\left(\frac{1}{q_{m,k}^0\,\Delta_{m,k}^2}\,\log\left(\frac{K}{\delta\Delta_{m,k}}\right)\right), \tag{14}$$

$$\sum_{m=1}^{M} \beta_m = \sum_{m=1}^{M} O\left(\max_{k \in \mathcal{A}_m} \frac{1}{q_{m,k}^0\,\Delta_{m,k}^2}\log\left(\frac{K}{\delta\Delta_{m,k}}\right)\right). \tag{15}$$

It is easy to see that the high-probability, instance-dependent upper bound on the stopping time of BBSEA given by (15) matches with the instance-dependent lower bound in (13) order-wise in δ and the instance-specific parameters $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$. These bounds are hence tight up to multiplicative factors.

V. Concluding Remarks

We studied best arm identification in a multi-armed bandit when the learner's access to the arms is via exploration intermediaries that we refer to as boxes, and the arm access probabilities of the boxes are unknown to the learner. The key challenge we addressed in the analysis is the non-uniqueness of optimal allocations in the information theoretic lower bound. We showed that by tracking the empirical average of *arbitrarily chosen* optimal allocations corresponding to running estimates of the problem instance, asymptotic optimality can be achieved. An interesting direction for future work is to develop efficient algorithms that admit non-asymptotic upper bounds; in this paper, we only do this in the special case where the arms are partitioned across the boxes. The main challenge here is to utilize estimates of arm selection probabilities of the different boxes to guide box selection. Given that any particular arm may be accessible via multiple boxes (with potentially different probabilities), it is not clear at the outset which box must be selected to maximise the chances of pulling a given arm, and therefore algorithms like LUCB do not generalize directly to our setting.

Acknowledgements: Kota Srinivas Reddy was supported by the Department of Science and Technology (DST), Govt. of India, through the INSPIRE faculty fellowship.
P. N. Karthik was supported by the National University of Singapore via grant A-0005994-01-00.
Nikhil Karamchandani and Jayakrishnan Nair acknowledge support from DST via grant CRG/2021/002923 and two SERB MATRICS grants.

REFERENCES

[1] K. S. Reddy, P. N. Karthik, N. Karamchandani, and J. Nair, "Best arm identification in bandits with limited precision sampling," *arXiv preprint arXiv:2305.06082*, 2023.

[2] P. N. Karthik and R. Sundaresan, "Detecting an odd restless Markov arm with a trembling hand," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5230–5258, 2021.

[3] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "Pac subset selection in stochastic multi-armed bandits." 2012.

[4] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[5] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*. PMLR, 2016, pp. 998–1027.

[6] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of machine learning research*, vol. 7, no. Jun, pp. 1079–1105, 2006.

[7] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2014, pp. 1–6.

[8] Y. Jedra and A. Proutiere, "Optimal best-arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 007–10 017, 2020.

[9] Z. Chen, P. N. Karthik, V. Y. F. Tan, and Y. M. Chee, "Federated best arm identification with heterogeneous clients," *arXiv preprint arXiv:2210.07780*, 2022.

[10] J. Scarlett, I. Bogunovic, and V. Cevher, "Overlapping multi-bandit best arm identification," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 2544–2548.

[11] S. A. Jain, S. Goenka, D. Bapna, N. Karamchandani, and J. Nair, "Sequential community mode estimation," *Performance Evaluation*, vol. 152, p. 102247, 2021.

[12] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, p. 4–22, mar 1985. [Online]. Available: https://doi.org/10.1016/0196-8858(85)90002-8