Best Arm Identification with Arm Erasures

Kota Srinivas Reddy
Indian Institute of Technology Madras
ksreddy@ee.iitm.ac.in

P. N. Karthik
Indian Institute of Technology Hyderabad
pnkarthik@ai.iith.ac.in

Vincent Y. F. Tan
National University of Singapore
vtan@nus.edu.sg

***Abstract*—In this paper, we address the problem of best arm identification (BAI) with arm erasures in a multi-armed bandit setting with finitely many arms. A *learner* who seeks to identify the best arm—the arm with the largest mean reward—samples arms sequentially, one at each time instant, and communicates the sampled arm to an *agent* through an erasure channel with a known erasure probability $\varepsilon \in (0,1)$. The learner *does not* receive any erasure feedback, and hence does not know whether the transmitted arm was erased by the channel. In instances where erasure does not occur, and the transmitted arm is successfully received by the agent, the agent promptly pulls the received arm. On the contrary, when erasure occurs, we analyse the following two distinct scenarios: (a) the agent randomly selects an arm, and (b) the agent selects the most recent successfully received arm. We assume that the instantaneous reward from the pulled arm is available to the learner, whose objective is to find the best arm as quickly as possible, subject to an upper bound on the error probability. Given $\delta \in (0,1)$, we derive a problem-dependent lower bound on the expected stopping time of any algorithm whose error probability is within δ. We also propose two successive elimination algorithms for each of the aforementioned scenarios (a), (b), and provide upper bounds on their stopping times that hold with probability $1-\delta$. To our best knowledge, this is the first work on BAI with arm erasures.**

I. Introduction

In machine learning, multi-armed bandits constitute a popular mathematical model for decision-making under uncertainty. Introduced by Thompson [1] in the context of clinical trials, multi-armed bandits have now found widespread applications in various fields including communication systems, power systems, medical imaging, recommendation systems, etc., where they serve as a mathematical model for allocating a scarce resource to a pool of two or more options (metaphorically called *arms*). Upon selection, each arm is assumed to yield a random reward from an unknown reward distribution. The choice of which arm to select is based on the available information about the arms and a pre-determined optimization criterion. The most commonly studied optimization criterion is cumulative regret, which measures the difference between the expected reward obtained and the reward obtained by a hypothetical oracle that knows the true reward distributions of the arms. Minimising the cumulative regret leads naturally to an *exploration-vs-exploitation* dilemma, a delicate balance-of-sorts between the desire to choose arms with high expected rewards (exploitation) against the need to explore other arms to acquire better information discrimination (exploration). See [2] for a survey of works on regret minimisation.

Contrary to the theme of regret minimisation is the theme of *pure exploration*, wherein the goal is to explore the arms with the eventual aim of validating the truth of one or more hypotheses without committing to any arm prematurely. Pure exploration problems fall within the umbrella framework of active sequential hypothesis testing devised by Chernoff [3] and Albert [4], and are instances of optimal stopping problems in decision theory. A class of pure exploration problems studied extensively in the literature is that of best arm identification (BAI) wherein the goal is to find the best arm—the arm with the largest mean reward—as quickly and accurately as possible. BAI is typically studied under one of two complementary regimes: (a) the fixed-confidence regime, wherein given a pre-specified threshold on the error probability, the goal is to minimise the expected time required to find the best arm, and (b) the fixed-budget regime, wherein given an arms sampling budget, the goal is to minimise the probability of error in finding the best arm. This paper focuses on BAI in the fixed-confidence regime under an additional constraint of arm erasures that is described next.

A. Motivation and Problem Setup

Our study is motivated by real-world scenarios exemplified in [5] such as a human controller (the *learner*) seeking to maneuver a drone (the *agent*) to a discrete target location in space, or a medical imaging expert seeking to guide a micro/nanobot inside the human body to capture images of a desired organ. Treating the target locations as analogous to the best arm, we explore a multi-armed bandit model with finitely many arms, each producing independent and identically distributed rewards. The learner, in the pursuit of finding the best arm, sequentially samples arms over time and transmits the selections to the agent via an erasure channel characterized by a *known* erasure probability $\varepsilon \in (0,1)$. We assume that the learner does not receive any feedback about erasure occurrences (i.e., *no erasure feedback*). In instances without arm erasure, the agent simply pulls the arm received successfully from the learner. On the contrary, in instances of arm erasure, we study the following two distinct arm selection strategies of the agent: (a) the agent samples an arm uniformly at random, or (b) the agent pulls the most recent successfully received arm. The arm selection strategy of the agent is predetermined through mutual agreement between the learner and the agent, and hence known to the learner. At each time instant, the instantaneous reward generated from the pulled arm is revealed to both agent and learner. The learner aims to recover the best arm in the shortest possible time, subject to an upper bound on the error probability.

979-8-3503-8284-6/24/$31.00 ©2024 IEEE

B. Analytical Challenges and Contributions

The key analytical challenge is that because of the possibility of arm erasures, the instantaneous reward observed by the learner at any given time may not necessarily be generated from the learner's transmitted arm, unlike in the classical BAI setting without arm erasures. For the case when the agent samples arms uniformly at random in instances of arm erasure, we derive a problem instance-dependent lower bound on the expected stopping time of any policy whose error probability is within a prescribed threshold, say $\delta \in (0,1)$. Augmenting the lower bound, we introduce two successive elimination algorithms, one for each of agent's arm sampling schemes in instances of arm erasure, and derive corresponding upper bounds on their stopping times that hold with probability $1-\delta$. In the case when the agent pulls the most recent successfully received arm in instances of arm erasure, the key idea in our algorithm is to repeatedly transmit the learner's selected arm, and to disregard the initial α fraction of rewards obtained, where α is a parameter of the algorithm. We show that the upper and the lower bounds match order-wise in the instance-dependent constants for the case of uniform arm selection by the agent. To our best knowledge, this is the first work on BAI in the presence of arm erasures.

C. Related Works

The topic of BAI has been investigated extensively in the literature, both in the fixed-budget and fixed-confidence regimes. Algorithms for fixed-confidence BAI are typically based on one of the following two themes: elimination-type algorithms with high-probability upper bounds on their stopping times, and tracking-type algorithms with upper bounds on their expected stopping times. Even-Dar et al. [6] proposed the earliest known elimination-type algorithm, called action elimination or successive elimination, which progressively eliminates sub-optimal arms based on upper and lower confidence bounds. The popular LUCB algorithm of [7] is also along similar lines as the elimination algorithm of [6]. Garivier and Kaufmann [8] obtained an upper bound on the expected stopping time of an algorithm that *tracks* the optimal proportions of arm pulls, and therefore matches with the lower bound in the asymptotic regime of vanishing error probabilities. Many subsequent works on fixed-confidence BAI have applied the core ideas in [6], [8] to several interesting problem settings such as linear bandits [9], federated learning [10], [11], transfer learning [12], bandits with Markov rewards from arms [13]–[15], and best policy identification in Markov decision processes [16], [17]. On the topic of fixed-budget BAI, the paper [18] obtains a minimax lower bound on the error probability. The authors therein show that their bound is order-wise tight in the exponent of the error probability. Yang and Tan [19] investigate fixed-budget BAI for linear bandits and propose an algorithm based on the idea of *G-optimal design*. They prove a minimax lower bound on the error probability and obtain an upper bound on the error probability of their algorithm, tailored to the setting of linear bandits, called OD-LINBAI. While the aforementioned works focus on scenarios without arm erasures, our work specifically addresses situations involving arm erasures when no erasure feedback is available to the learner. The more straightforward scenario, where erasure feedback is indeed available to the learner, is essentially a specific case of a more general problem setting studied in [20].

The model of arm erasures studied in this paper is inspired from the paper [5] that is based on the theme of regret minimisation; on the contrary, we study the problem of fixed-confidence BAI. As such, it is well known [21], [22] that algorithms which work well within the framework of regret minimisation, do not necessarily work well within the framework of BAI, and vice-versa. For connections of the current problem setting to the adversarial bandit setting, see [5].

II. PRELIMINARIES

We write $\mathbb{N}$ to denote the set of positive integers. Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$. We consider a multi-armed bandit with $K \geq 2$ arms. Arm $a \in [K]$ is associated with a reward distribution $\nu_a = \mathcal{N}(\mu_a, 1)$, where $\mathcal{N}(\mu, 1)$ denotes a Gaussian distribution with mean μ and unit variance. Let $\boldsymbol{\mu} := [\mu_1, \ldots, \mu_K]^\top \in \mathbb{R}^K$. In the sequel, we refer to $\boldsymbol{\mu} \in \mathbb{R}^K$ as the *problem instance*. Let $a^\star(\boldsymbol{\mu}) = \arg\max_{a \in [K]} \mu_a$ denote the *best* arm under the instance $\boldsymbol{\mu}$. A learner who does not have prior knowledge of $\boldsymbol{\mu}$, wishes to identify the best arm by sequentially sampling the arms, one at each time instant $t \in \mathbb{N}_0$. Let A_t denote the arm selected by the learner at time t. The learner transmits its selection A_t to an agent located far away over an erasure channel with a *fixed* and *known* erasure probability $\varepsilon \in (0,1)$. In other words, with probability $1-\varepsilon$, the transmitted arm A_t is successfully received by the agent, and with probability ε, the transmitted arm is erased by the channel. Let $\tilde{B}_t \in \{A_t, \text{null}\}$ denote the channel output at time t; here, $\tilde{B}_t = \text{null}$ if A_t is erased by the channel.

We assume that the erasures are *independent* across time and also independent of the arm selections of the learner. Further, we assume that *no erasure feedback* is available to the learner, i.e., the learner does not observe $\tilde{B}_t$ and hence does not know whether the transmitted arm A_t was successfully received by the agent. If $\tilde{B}_t = A_t$, then the agent pulls arm A_t. On the contrary, if $\tilde{B}_t = \text{null}$, we analyse two distinct scenarios: (a) the agent pulls arm an arm uniformly at random, or (b) the agent pulls the most recent successfully received arm from the learner. In the latter scenario, denoting $\tilde{A}_t$ as the arm pulled by the agent at time t, we have

$$\tilde{A}_t = A_t \, \mathbf{1}\{\tilde{B}_t = A_t\} + \tilde{A}_{t-1} \, \mathbf{1}\{\tilde{B}_t = \text{null}\}. \tag{1}$$

Without loss of generality, we assume that if an erasure occurs at time $t = 0$ (i.e., $\tilde{B}_0 = \text{null}$), then $\tilde{A}_0 \sim \text{Unif}([K])$, where $\text{Unif}([K])$ denotes the uniform distribution on $[K]$. Let X_t denote the reward generated from pulling arm $\tilde{A}_t$ at time t. The reward X_t is revealed to both agent and learner. Let $(A_{0:t}, X_{0:t}) := (A_0, X_0, \ldots, A_t, X_t)$ denote the history of arm selections and observations available to the learner at time t.

We note here that X_t may not necessarily be the reward from arm A_t.

Let $\pi = \{\pi_t\}_{t=0}^{\infty}$ denote any generic best arm identification *policy* (or algorithm) of the learner, where for every $t \in \mathbb{N}_0$, the function π_t maps the history $(A_{0:t-1}, X_{0:t-1})$ to one of the following actions:

- Select arm A_t according to a deterministic or randomised rule.
- Stop and declare the estimated best arm.

Let τ_π and $\widehat{a}_\pi$ denote respectively the (random) stopping time and the estimate of the best arm under π. For each $\delta \in (0,1)$, our interest is in the class of δ-*probably correct* (δ-PC) algorithms defined by

$$\Pi(\delta) \coloneqq \{\pi : \mathbb{P}_{\boldsymbol{\mu}}(\tau_\pi < +\infty) = 1, \\ \mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_\pi \neq a^\star(\boldsymbol{\mu})) \leq \delta \quad \forall \boldsymbol{\mu}\}. \quad (2)$$

Here, and throughout the paper, we write[1] $\mathbb{P}_{\boldsymbol{\mu}}$ and $\mathbb{E}_{\boldsymbol{\mu}}$ to denote probabilities and expectations under the instance $\boldsymbol{\mu}$. In this paper, we design δ-PC algorithms and benchmark their stopping times against information theoretic lower bounds.

For the remainder of the paper, we fix the underlying instance $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$, and assume without loss of generality that $\mu_1 > \mu_2 > \cdots > \mu_K$. All the results presented in this paper are under this assumption.

III. Uniform Arm Sampling by Agent Under Erasure Instances

In this section, we outline our results for the scenario when the agent samples arms uniformly at random on every occurrence of arm erasure, and when no arm erasure occurs, the agent simply pulls the arm received from the learner. Under this arm sampling strategy of the agent (which is also known to the learner), we note that almost surely,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}}[X_t | A_t = a] \\
&= (1-\varepsilon)\, \mathbb{E}_{\boldsymbol{\mu}}[X_t | A_t = a,\, \mathcal{E}_t^c] + \varepsilon\, \mathbb{E}_{\boldsymbol{\mu}}[X_t | A_t = a,\, \mathcal{E}_t] \\
&= (1-\varepsilon)\, \mu_a + \frac{\varepsilon}{K} \sum_{a'=1}^{K} \mu_{a'} \\
&= (1-\varepsilon)\, \mu_a + \varepsilon\, \bar{\mu}, \qquad (3)
\end{aligned}$$

where $\mathcal{E}_t$ denotes the event that an erasure occurred at time t, and $\bar{\mu} = \sum_{a'=1}^{K} \mu_{a'}/K$ denotes the average of arm means under the instance $\boldsymbol{\mu}$. Eq. (3) reveals that the conditional average reward of every arm is simply shifted and scaled when the agent samples arms uniformly under erasure events. Letting μ_a^ε denote the quantity on the right-hand side of (3), we note that the best arm under the instance $\boldsymbol{\mu}^\varepsilon \coloneqq [\mu_a^\varepsilon : a \in [K]]^\top$ is identical to that under $\boldsymbol{\mu}$. Additionally, (3) reveals that intuitively, the complexity of BAI with arm erasures for the instance $\boldsymbol{\mu}$ should be equal to that without erasures for the instance $\boldsymbol{\mu}^\varepsilon$. We show that this is indeed true, as formalised in the below result.

Proposition 1: Given any $\delta \in (0,1)$ and a δ-PC algorithm $\pi \in \Pi(\delta)$, we have

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\pi] \geq \log\left(\frac{1}{4\delta}\right) \cdot \left\{\sum_{a=1}^{K} \frac{1}{\Delta_{\varepsilon,a}^2/2}\right\}, \qquad (4)$$

where $\Delta_{\varepsilon,a} \coloneqq (1-\varepsilon)\,\Delta_a$, with $\Delta_a \coloneqq \mu_1 - \mu_a$ for $a \neq 1$, and $\Delta_1 \coloneqq \Delta_2$.

Notice that $\Delta_{\varepsilon,a} = \mu_1^\varepsilon - \mu_a^\varepsilon$ for all $a \neq 1$, and $\Delta_{\varepsilon,1} = \mu_1^\varepsilon - \mu_2^\varepsilon$. Thus, the quantity $\sum_{a=1}^{K} 1/(\Delta_{\varepsilon,a}^2/2)$ corresponds to the complexity of identifying the best arm for the instance $\boldsymbol{\mu}^\varepsilon$. Given any $\delta \in (0,1)$, Proposition 1 delineates the minimum expected number of arm selections by the learner required to find the best arm with error probability not exceeding δ. Its proof is based on standard change-of-measure arguments for multi-armed bandits.

In the next section, we provide an algorithm based on the principle of successive elimination [6] whose stopping time satisfies a high-probability upper bound that almost matches the lower bound in (4) order-wise in the instance-specific constants.

A. Successive Elimination for BAI with Arm Erasures and Uniform Arm Sampling by Agent Under Erasures

Before we present the algorithm, we first introduce some algorithm-specific notations. Our proposed algorithm operates in rounds. We use n to denote the round number and t to denote the running time. For any arm a, we write $t_a(n)$ to denote the number of times arm a is pulled up to round n. We denote the empirical mean of arm a after $t_a(n)$ pulls by $\widehat{\mu}_a(n)$. We write S to denote the set of active arms (i.e., candidate best arms) in any given round. Given $\delta \in (0,1)$, we define $\alpha_\delta(x) \coloneqq \sqrt{\frac{2\log(8\,K\,x^2/\delta)}{x}}$. The upper and lower confidence bounds on $\widehat{\mu}_a(n)$, with a confidence interval of length $\alpha_\delta(t_a(n))$, are denoted as $\mathrm{UCB}_a(n)$ and $\mathrm{LCB}_a(n)$, respectively, and defined via

$$\mathrm{UCB}_a(n) \coloneqq \widehat{\mu}_a(n) + \alpha_\delta(t_a(n)), \qquad (5)$$

$$\mathrm{LCB}_a(n) \coloneqq \widehat{\mu}_a(n) - \alpha_\delta(t_a(n)). \qquad (6)$$

Our algorithm for BAI under erasures and uniform arm sampling by the agent under erasure events, termed *Successive Elimination with Uniform arm pulls by agent under erasures*, or SEUNIF in short, maintains a list of active arms (those that are in contention for being the best arm) and progressively eliminates sub-optimal arms from the list. Arm a is designated to be sub-optimal (and hence eliminated) in round n if there exists an arm a' such that $\mathrm{LCB}_{a'}(n) > \mathrm{UCB}_a(n)$. The elimination process continues until only a single arm is left, at which point the algorithm stops and declares the single remaining arm as the best arm. The pseudocode for SEUNIF is presented in Algorithm 1.

[1]To be precise, it is instructive to write $\mathbb{P}_{\boldsymbol{\mu}}^\pi$ and $\mathbb{E}_{\boldsymbol{\mu}}^\pi$, explicitly denoting the dependence of the preceding quantities on the underlying policy π. However, for notational brevity, we drop the superscript π and urge the reader to bear this dependence in mind.

Algorithm 1 Successive Elimination with Uniform arm pulls by agent under erasures – SEUNIF

Input: $K \in \mathbb{N}$, $\delta \in (0,1)$
Output: $\widehat{a}_{\text{SEUNIF}} \in [K]$ (best arm)
Initialization: $n = 0$, $S = [K]$
while $|S| > 1$ **do**
  $n \leftarrow n + 1$
  Pull each arm $a \in S$ once.
  Set $t_a(n) \leftarrow t_a(n-1) + 1$. Update $\widehat{\mu}_a(n), \text{UCB}_a(n)$ and $\text{LCB}_a(n)$ for all $a \in S$.
  if $\exists a' \in S$ such that $\text{UCB}_a(n) < \text{LCB}_{a'}(n)$ **then**
    $S \leftarrow S \setminus \{a\}$
  end if
  if $|S| = 1$ **then**
    $\widehat{a}_{\text{SEUNIF}} \leftarrow a \in S$
  end if
end while
return $\widehat{a}_{\text{SEUNIF}}$.

B. Performance Characterisation

In this section, we characterise the performance of SEUNIF. Notice that SEUNIF does not exploit the knowledge of the channel erasure probability ε. Nevertheless, it achieves almost optimal performance as exemplified by the below result.

Theorem 2: For any $\delta \in (0,1)$, the following hold with probability greater than $1-\delta$.

1) SEUNIF outputs the best arm correctly.
2) The stopping time of SEUNIF is upper bounded by $\sum_{a=1}^{K} T_a$, where $T_a := 1 + \frac{102}{\Delta_{\varepsilon,a}^2} \log\left(\frac{64\sqrt{\frac{8K}{\delta}}}{\Delta_{\varepsilon,a}^2}\right) = \tilde{O}\left(\frac{1}{\Delta_{\varepsilon,a}^2}\right)$, and $\tilde{O}(\cdot)$ hides the dependence on log factors.

Notice that the high-probability upper bound of $\sum_{a=1}^{K} T_a$ in Theorem 2 matches the lower bound in (4) order-wise (up to logarithmic factors) in the instance-specific constants and the erasure probability ε. The proof of Theorem 2 follows along the lines of the proof of [6, Theorem 8] by noting that the complexity of BAI for the instance $\boldsymbol{\mu}$ with arm erasures and uniform arm sampling by agent under erasure events (occurring with probability ε), is identical to that of BAI without erasures for the modified instance $\boldsymbol{\mu}^{\varepsilon}$.

IV. Most Recent Successfully Received Arm Sampling by Agent Under Erasures

In this section, we investigate the more challenging case when the agent pulls the most recent successfully received arm under erasure events, i.e., the agent continues to pull the previously pulled arm under erasure events. The primary challenge here is that the agent's arm sampling scheme introduces *memory* into the system. To see this, we note here that under the instance $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^\top$, almost surely,

$$\mathbb{E}_{\boldsymbol{\mu}}[X_t | A_{0:t}, X_{0:t-1}] = (1-\varepsilon) \sum_{u=0}^{t} \varepsilon^{t-u} \mu_{A_u} + \frac{\varepsilon^{t+1}}{K} \sum_{a'=1}^{K} \mu_{a'}. \tag{7}$$

Algorithm 2 Modified Successive Elimination Algorithm when agent pulls the previous arm under erasures – MSEA

Input: $K \in \mathbb{N}$, $\delta \in (0,1)$, $\alpha \in (0,1]$
Output: $\widehat{a}_{\text{MSEA}} \in [K]$ (best arm)
Initialization: $n = 0$, $t = 0$, $S = [K]$
while $|S| > 1$ **do**
  $n \leftarrow n + 1$
  Pull each active arm $a \in S$ for a total of t_n times, and ignore the first $\lfloor \alpha t_n \rfloor$ pulls and the associated rewards.
  For all $a \in [K]$, set $t_a(n) \leftarrow t_a(n-1) + \lceil (1-\alpha)\, t_n \rceil$. Also, update $\widehat{\mu}_a(n), \text{UCB}_a(n)$ and $\text{LCB}_a(n)$ based on the last $\lceil (1-\alpha)\, t_n \rceil$ rewards seen from arm a.
  if $\exists a' \in S$ such that $\text{UCB}_a(n) < \text{LCB}_{a'}(n)$ **then**
    $S \leftarrow S \setminus \{a\}$.
  end if
  if $|S| = 1$ **then**
    $\widehat{a}_{\text{MSEA}} \leftarrow a \in S$
  end if
end while
return $\widehat{a}_{\text{MSEA}}$.

The second term on the right-hand side of (7) is an artefact of the assumption that the agent samples arm uniformly at random if an erasure occurs at time $t = 0$. Observe that the first term on the right-hand side of (7) is a function of $A_{0:t}$, the history of all the arm selections of the learner up to time t, unlike in the case of uniform arm sampling by agent under erasure events (where the analogous first term is a function of the most recent arm A_t only; see (3)). Because of the more complicated dependence of (7) on all previous arm selections of the learner, it is not straightforward to derive the complexity of finding the best arm under any δ-PC policy. While we leave this task open for future exploration, we provide below a modified successive elimination algorithm for BAI that operates in the presence of memory induced by the agent's strategy.

A. Modified Successive Elimination Algorithm

We adapt the SEUNIF algorithm from before, with an additional ingredient of repeated pulling of arms. Our algorithm, termed *Modified Successive Elimination Algorithm* (or MSEA in short), takes as input the following parameters: $K \in \mathbb{N}$, $\delta \in (0,1)$, and $\alpha \in (0,1]$. See Algorithm 2 for the pseudocode. The operation of MSEA is as follows. In each round n, the algorithm transmits each active arm to the agent a certain pre-determined number of times, say t_n, and observes the associated rewards. Out of the total t_n arm transmissions and the associated rewards, the algorithm disregards the initial α fraction of the transmissions and rewards. Writing $t_a^{\text{eff}}(n-1)$ (resp. $t_a(n-1)$) to denote the *effective* (resp. total) number of times arm active arm a is transmitted until the end of round $n-1$, we have the relations $t_a^{\text{eff}}(n) = t_a^{\text{eff}}(n-1) + \lceil (1-\alpha)\, t_n \rceil$ and $t_a(n) = t_a(n-1) + t_n$ for all active arms a. Using only the final $\lceil (1-\alpha)\, t_n \rceil$ fraction of rewards to compute/update

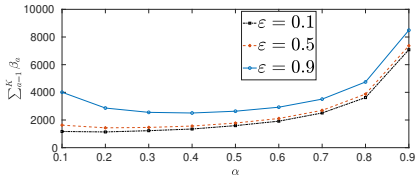

Figure 1. Plot of the upper bound of Theorem 3 as a function of α for varying values of erasure probability ε, for a sample instance with $K = 5$, means $\mu_1 = 1$ and $\mu_a = -1$ for all $a \neq 1$, and $\delta = 0.3$.

the empirical mean of each active arm, the algorithm employs the modified relations

$$\mathrm{UCB}_a(n) := \widehat{\mu}_a(n) + \alpha_\delta(t_a^{\mathrm{eff}}(n)), \tag{8}$$

$$\mathrm{LCB}_a(n) := \widehat{\mu}_a(n) - \alpha_\delta(t_a^{\mathrm{eff}}(n)), \tag{9}$$

to compute the upper/lower confidence bounds for each $a \in S$ and to eliminate the sub-optimal arms in round n. By carefully designing t_n, we give theoretical guarantees on the performance of MSEA in the next section.

B. *Performance of* MSEA

Let $T := \max\left\{\left\lceil \frac{\log(\frac{2K}{\delta}+1)}{\alpha \log(1/\varepsilon)} \right\rceil, 1\right\}$, and let $t_n := nT$.

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

1) MSEA outputs the best arm correctly.
2) The stopping time of MSEA is upper bounded by $\sum_{a=1}^{K} \beta_a$, where $\beta_a := \frac{T'_a}{1-\alpha} + \sqrt{\frac{2\,T\,T'_a}{1-\alpha}} + T$, $T'_a := 1 + \frac{102}{\Delta_a^2} \log\left(\frac{64\sqrt{\frac{8K}{\delta}}}{\Delta_a^2}\right)$, $\Delta_a := \mu_1 - \mu_a$ for all $a \neq 1$, and $\Delta_1 := \mu_1 - \mu_2$.

We note that the quantities T and $T'_a/(1-\alpha)$ exhibit opposing behaviour with respect to the parameter α: as α increases, the former decreases, while the latter increases. This inherent duality gives rise to a trade-off, the overall effect of which manifests in the upper bound expression $\sum_{a=1}^{K} \beta_a$. Intuitively, then, one anticipates the existence of an optimal α value that minimizes the upper bound. However, determining this optimal value apriori is infeasible as it is a function of the underlying problem instance that is unknown to the algorithm. Figure 1 shows a plot of $\sum_{a=1}^{K} \beta_a$ versus α with varying values of erasure probability ε, for a sample instance with $K = 5$, means $\mu_1 = 1$ and $\mu_a = -1$ $\forall a \neq 1$, and $\delta = 0.3$.

V. Discussion and Future Work

The inquisitive reader may ponder which among the two strategies of pulling arms uniformly at random and pulling the previous arm under erasure instances, will result in a smaller stopping time. To partly answer this query, one may investigate which of the two upper bounds presented in Theorems 2 and 3 is smaller. While the order-wise match (up to logarithmic factors) between the upper bound in Theorem 2 and the lower bound in Proposition 1 is established for the case of uniform sampling, a commensurate assessment for the upper bound in Theorem 3 remains unaddressed in our study due to the lack of a corresponding lower bound for comparison. Consequently, our current study precludes any concrete comparison of the relative magnitudes of the upper bounds on the stopping time coming from Theorems 2 and 3.

Nevertheless, consider a modified scenario where the agent adopts the strategy of pulling the previous arm until it observes R consecutive erasures, transitioning to uniformly sampling the next arm on the occurrence of the $(R+1)$-th erasure, for a fixed $R \in \mathbb{N}_0$. In this context, it is notable that under the instance $\boldsymbol{\mu}$, the conditional expectation $\mathbb{E}_{\boldsymbol{\mu}}[X_t | A_{0:t}, X_{0:t-1}]$ takes the form

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}}[X_t | A_{0:t}, X_{0:t-1}] \\ &= (1-\varepsilon) \sum_{u=t-R}^{t} \varepsilon^{t-u}\, \mu_{A_u} + \frac{\varepsilon^{R+1}}{K} \sum_{a'=1}^{K} \mu_{a'}.\end{aligned} \tag{10}$$

Notably, (10) reduces to (3) when $R = 0$, while the substitution $R = R_t = t$ recovers (7). This formulation delineates a spectrum of arm sampling strategies parameterized by R, ranging from uniform arm selection ($R = 0$) to persistently pulling the same arm (with $R = R_t \to \infty$ as $t \to \infty$). The conditional expectation term in (10), incorporating dependencies on arms $A_t, A_{t-1}, \ldots, A_{t-R}$, underscores that a larger value of R provides the learner with more information from past actions that can potentially be exploited to find the best arm sooner. We therefore anticipate that the agent's strategy to consistently pull the same arm yields the smallest stopping time within this spectrum of strategies. Formalizing this intuition presents a promising avenue for future research.

While the results of this paper are for a fixed error probability threshold $\delta \in (0,1)$, exploring the asymptotics of the problem as $\delta \downarrow 0$, akin to [8], would be an intriguing avenue for future work. This would entail deriving an asymptotic lower bound on the stopping time and developing an algorithm along the lines of the well-known Track-and-Stop algorithm [8] that potentially matches the lower bound in performance. Notably, the asymptotic analysis of the modified scenario of R consecutive erasures introduced above presents many analytical challenges, the foremost being the careful consideration of the previous $R+1$ arm selections of the learner at any given time. Demonstrating that the asymptotic problem complexity for R consecutive erasures interpolates neatly between those of uniform arm pulling and pulling the same arm would be particularly insightful.

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 cumulative professional development allowance (CPDA), courtesy Govt. of India. Vincent Y. F. Tan was supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018) and the Singapore Ministry of Education Academic Research Fund Tier 2 under grant number A-8000423-00-00.

REFERENCES

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[2] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[3] H. Chernoff, "Sequential design of experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755–770, 1959.

[4] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *The Annals of Mathematical Statistics*, vol. 32, no. 3, pp. 774–799, 1961.

[5] O. A. Hanna, M. Karakas, L. F. Yang, and C. Fragouli, "Multi-arm bandits over action erasure channels," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 1312–1317.

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

[9] 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.

[10] K. S. Reddy, P. N. Karthik, and V. Y. F. Tan, "Almost cost-free communication in federated best arm identification," *arXiv preprint arXiv:2208.09215*, 2022.

[11] 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.

[12] O. Neopane, A. Ramdas, and A. Singh, "Best arm identification under additive transfer bandits," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021, pp. 464–470.

[13] V. Moulos, "Optimal best markovian arm identification with fixed confidence," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] P. N. Karthik, K. S. Reddy, and V. Y. F. Tan, "Best arm identification in restless Markov multi–armed bandits," *IEEE Transactions on Information Theory*, 2022.

[15] P. N. Karthik, V. Y. F. Tan, A. Mukherjee, and A. Tajer, "Optimal best arm identification with fixed confidence in restless bandits," *arXiv preprint arXiv:2310.13393*, 2023.

[16] J. Taupin, Y. Jedra, and A. Proutiere, "Best policy identification in linear mdps," in *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2023, pp. 1–8.

[17] A. Al Marjani, A. Garivier, and A. Proutiere, "Navigating to the best policy in markov decision processes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 852–25 864, 2021.

[18] A. Carpentier and A. Locatelli, "Tight (lower) bounds for the fixed budget best arm identification bandit problem," in *Conference on Learning Theory*. PMLR, 2016, pp. 590–604.

[19] J. Yang and V. Y. F. Tan, "Minimax optimal fixed-budget best arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 253–12 266, 2022.

[20] 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.

[21] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in finitely-armed and continuous-armed bandits," *Theoretical Computer Science*, vol. 412, no. 19, pp. 1832–1852, 2011.

[22] Z. Zhong, W. C. Cheung, and V. Y. F. Tan, "Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits," *arXiv preprint arXiv:2110.08627*, 2021.

[23] 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.