Best Restless Markov Arm Identification

Karthik Periyapattana Narayana Prasad, Srinivas Reddy Kota, and Vincent Y. F. Tan
National University of Singapore
Email: {karthik, ksreddy, vtan}@nus.edu.sg

***Abstract*—We study the problem of best arm identification in multi-armed bandits when each arm is an ergodic Markov process that evolves whether or not the arm is selected (*restless* arms). The evolution of each arm's Markov process is governed by its transition probability matrix (TPM). A decision entity that knows the set of arm TPMs but not the exact mapping of the TPMs to the arms, wishes to find the index of the best arm as quickly as possible, subject to an upper bound on the error probability. We derive an asymptotic lower bound on the expected time required to find the best arm, where the asymptotics is as the error probability vanishes. Also, we design a policy that, for an input parameter R, forcibly selects an arm that has not been selected for R consecutive time instants, and achieves an upper bound that is monotonically non-increasing in R. Showing that, in general, the lower bound and the limiting value of the upper bounds as $R \to \infty$ match, appears to be difficult and remains open. These bounds are, however, shown to match in the special case when the TPM of each arm has identical rows, i.e., the arms yield independent and identically distributed observations.**

I. Introduction

This paper studies the problem of finding the best arm in a multi-armed bandit when (a) each arm is a Markov process, and (b) the arms are *restless*. Here, the phrase 'restless' means that the Markov process of each arm continues to evolve whether or not the arm is selected. We assume that (a) the arms are independent of one another, and (b) the Markov process of each arm evolves on a common, finite state space $\mathcal{S}$ according to a rule prescribed by the arm's transition probability matrix (TPM). Given $\epsilon > 0$, a decision entity that knows the set of arm TPMs but not the exact mapping of the TPMs to the arms, wishes to find the index of the best arm as quickly as possible while keeping its error probability within ϵ. Our interest is to analyse the growth rate of the expected time required to find the best arm in the asymptotic regime as $\epsilon \downarrow 0$.

A. Related Work

The above described problem falls within the class of optimal stopping problems in decision theory, and can be embedded within the framework of active sequential hypothesis testing studied by Chernoff [1] and Albert [2]. Prior works on best arm identification have dealt with (a) independent and identically distributed (*i.i.d.*) observations from the arms [3], [4], and (b) rested Markov arms [5], where the unselected arms do not evolve and remain frozen. We extend the results of these works to the more difficult setting of restless Markov arms. For a related problem of odd arm identification in restless Markov multi-armed bandits, see [6], [7].

An examination of the results in [3]–[5] reveals that the growth rate of the expected time required to find the best arm with an error probability no more than ϵ is of the order $\Theta(\log(1/\epsilon))$. We anticipate that a similar growth rate holds in the setting of restless Markov arms studied in this paper.

B. Key Challenges to Overcome in the Analysis

The continued evolution of the arms necessitates the decision entity to maintain, for each of the arms, a record of (a) the time elapsed since an arm was last selected (the arm's *delay*), and (b) the Markovian state of an arm observed at the time of its most recent selection (the arm's *last observed state*). As noted in [6], the arm delays are integer-valued and cannot be handled on a machine with finite memory. As a result, any policy that selects the arms conditionally based on the values of the arm delays and the last observed states, as is typical of policies in the setting of restless arms (see [6] for a discussion on this), is not practically implementable. These issues do not arise in the prior works [3]–[5] because the arm delays are superfluous in the settings of these works.

In [6], the authors demonstrate that the arm delays and the last observed states together form a controlled Markov process whose state space is countably infinite. Under a certain *trembling hand* model for selecting the arms, they show that the controlled Markov process is ergodic with respect to every stationary control (or arm selection) policy. It is this ergodicity property that is pivotal to the converse and the achievability analyses in [6]. However, it is not clear if such an ergodicity property holds when the trembling hand assumption is relaxed as in this paper.

C. Our Contributions

We show that the expected time required to find the best arm with an error probability no more than ϵ is lower bounded by a term of the order $O(\log(1/\epsilon))$. We obtain the expression for the problem instance-dependent constant multiplying $\log(1/\epsilon)$ in the lower bound in terms of the arm TPMs. Furthermore, we devise a policy that, for an input parameter R, forcibly selects an arm that has not been selected for R consecutive time instants, and show that the expected time required by this policy to find the best arm with an error probability no more than ϵ is upper bounded by a term of the order $\Omega(\log(1/\epsilon))$. We obtain the expression for the constant multiplying $\log(1/\epsilon)$ in the upper bound and show that it is monotonically non-increasing in R. Showing that, in general, the lower bound and the limiting value of the upper bounds as $R \to \infty$ match, appears to be a difficult problem and remains open. Notwithstanding this, we show that these bounds match in the special case when the TPM of each arm has identical rows, i.e.,

978-1-6654-8341-4/22/$31.00 ©2022 IEEE

each arm yields *i.i.d.* observations. Thus, our results specialise to the results of [3], [4] when the *i.i.d.* observations of each arm come from a finite alphabet. In contrast to the policies of [6], [7], our policy is practically implementable for all values of R. Our analyses of the lower and the upper bounds can be extended to more general sequential hypothesis testing problems such as finding the exact mapping of the TPMs to the arms, finding the second-best arm, etc.

In the remainder of this paper, we present a summary of our findings. The detailed proofs of the results stated here may be found in [8].

II. Notations and Preliminaries

We consider a multi-armed bandit with $K \geq 2$ arms and let $\mathcal{A} = \{1, \ldots, K\}$ denote the set of arms. Each arm is associated with a time-homogeneous and ergodic discrete-time Markov process evolving on a common, finite state space $\mathcal{S}$. The state transitions on each arm are governed by the arm's transition probability matrix (TPM). Given TPMs $P_1, \ldots, P_K$ and a permutation (bijective function) $\sigma : \mathcal{A} \to \mathcal{A}$, let $C = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$ denote a problem instance in which the TPM assigned to arm a is $P_{\sigma(a)}$, $a \in \mathcal{A}$. In the sequel, we refer to C as an *assignment of the TPMs*.

Let μ_k denote the unique stationary distribution of the TPM P_k, $k \in \{1, \ldots, K\}$. Given $f : \mathcal{S} \to \mathbb{R}$, the best arm

$$a^\star(C) \coloneqq \arg\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{S}} f(i)\, \mu_{\sigma(a)}(i). \tag{1}$$

That is, $a^\star(C)$ is the arm with the largest average value of f computed under the arm's stationary distribution. We assume that $a^\star(C)$ is unique, i.e., there is a single best arm. A problem instance C' is said to be *alternative* to the problem instance C if $a^\star(C') \neq a^\star(C)$, i.e., the indices of the best arm in C and C' are distinct. We write $\mathsf{Alt}(C)$ to denote the set of all problem instances alternative to C. A decision entity that knows $P_1, \ldots, P_K$, but does not know C, wishes to find $a^\star(C)$, the best arm index in C, as quickly as possible, subject to an upper bound on its error probability. To do so, the decision entity selects the arms sequentially, one at each discrete time instant $t \in \{0, 1, 2, \ldots\}$. Let $\{X_t^a : t \geq 0\}$ denote the Markov process of arm a, and let A_t be the arm selected at time t. The unselected arms continue to exhibit state transitions (*restless* arms). Upon selecting arm A_t, the decision entity observes the state $\bar{X}_t = X_t^{A_t}$ of arm A_t. Let $(A_{0:t}, \bar{X}_{0:t}) = (A_0, \bar{X}_0, \ldots, A_t, \bar{X}_t)$ denote the history of all the arm selections and observations seen up to time t. We assume that all random variables are defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A *policy* π of the decision entity is defined by a collection of functions $\{\pi_t : t \geq 0\}$ such that at each time $t \geq 0$, the function π_t does one of the following based on the history $(A_{0:t-1}, \bar{X}_{0:t-1})$:

- stop and declare the index of the best arm;
- choose to pull arm A_t according to a deterministic or a randomised rule.

We write π to denote a generic policy, $\tau(\pi)$ to denote the stopping time of policy π, and $\theta(\tau(\pi))$ to denote the index of the best arm declared by the policy π at the stopping time. We write P_C^π and $\mathbb{E}_C^\pi$ to denote probabilities and expectations under an assignment of the TPMs C and a policy π.

Given an error probability threshold $\epsilon > 0$, let $\Pi(\epsilon)$ denote the collection of all policies whose error probability at the stopping time is no more than ϵ, i.e.,

$$\Pi(\epsilon) \coloneqq \{\pi : \ P_C^\pi(\theta(\tau(\pi)) \neq a^\star(C)) \leq \epsilon \ \forall\, C\}. \tag{2}$$

Because C is not known to the decision entity, the policies in $\Pi(\epsilon)$ must satisfy the criterion that their error probability is no more than ϵ for all possible C ($K!$ in total). Prior works [3]–[5] show that $\inf_{\pi \in \Pi(\epsilon)} \mathbb{E}_C^\pi[\tau(\pi)]$ grows as $\Theta(\log(1/\epsilon))$ in the limit as $\epsilon \downarrow 0$. We anticipate that a similar growth rate holds in the setting of restless Markov arms. Our interest is in characterising, or at least bounding, the value of

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)}. \tag{3}$$

For simplicity, we assume that every policy starts by selecting arm 1 at time $t = 0$, arm 2 at time $t = 1$, etc., and arm K at time $t = K - 1$. This ensures that the Markov process of each arm is observed at least once.

III. A Markov Decision Problem and Log-Likelihood Ratios

The continued evolution of the arms in our work necessitates the decision entity to maintain a record of (a) the time elapsed since an arm was last selected (the arm's *delay*), and (b) the Markovian state of an arm observed at the time of its most recent selection (the arm's *last observed state*). Let $d_a(t)$ and $i_a(t)$ denote respectively the delay and the last observed state of arm a at time t. Notice that $d_a(t)$ and $i_a(t)$ may be computed only when each arm is selected at least once, and are therefore defined for all $t \geq K$. Let $\underline{d}(t) \coloneqq (d_1(t), \ldots, d_K(t))$ and $\underline{i}(t) \coloneqq (i_1(t), \ldots, i_K(t))$. We shall adopt the following rule for updating $(\underline{d}(t), \underline{i}(t))$ to $((\underline{d}(t+1), \underline{i}(t+1)))$: if $A_t = a$, then $d_{a'}(t+1) = d_a(t) + 1$ for all $a' \neq a$, and $d_a(t+1) = 1$, indicating that arm a was sampled at time t. Additionally, $i_{a'}(t+1) = i_a$ for all $a' \neq a$, and $i_a = X_t^a$. Therefore, (a) the delay of each arm is positive integer valued, and (b) the delay of an unselected arm increments by 1 and its last observed state remains unchanged. Let $\mathbb{S}$ denote the set of all possible values of $(\underline{d}(t), \underline{i}(t))$ for $t \geq K$. Clearly, $\mathbb{S}$ is countably infinite.

As noted in [6, Section II], the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a controlled Markov process with $\{A_t : t \geq 0\}$ as the set of controls. We are thus in the setting of a Markov decision problem (MDP) whose state space is $\mathbb{S}$, action space is $\mathcal{A}$, the state at time t is $(\underline{d}(t), \underline{i}(t))$, and the action at time t is A_t. At any given time $t \geq K$, based on the history $\{A_{0:t-1}, \bar{X}_{0:t-1}\}$ (which is equivalently described by $\{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}$), the decision entity selects arm A_t and forms $(\underline{d}(t+1), \underline{i}(t+1))$. This continues until the stopping time.

Given any $(\underline{d}, \underline{i}) \in \mathbb{S}$, $a \in \mathcal{A}$, and $j \in \mathcal{S}$, let

$$N(n, \underline{d}, \underline{i}, a, j) \coloneqq \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t)=\underline{d},\, \underline{i}(t)=\underline{i},\, A_t=a,\, X_t^a=j\}}. \quad (4)$$

The quantity in (4) is the number of times up to time n the state $(\underline{d}, \underline{i})$ is observed, control (or arm) a is chosen subsequently, and state j is observed on arm a. Given C, C' and a policy π, we write $Z^{\pi}_{CC'}(n)$ to denote the log-likelihood ratio (LLR) of all the arm selections and observations seen up to time n under the assignment of the TPMs C versus that under C'. It can be shown that

$$Z^{\pi}_{CC'}(n) = \sum_{a=1}^{K} \log \frac{P^{\pi}_{C}(X^a_{a-1})}{P^{\pi}_{C'}(X^a_{a-1})} + \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \sum_{j\in\mathcal{S}} N(n, \underline{d}, \underline{i}, a, j)\ \log \frac{(P^a_C)^{d_a}(j|i_a)}{(P^a_{C'})^{d_a}(j|i_a)}. \quad (5)$$

In (5), P^a_C denotes the TPM of arm a under the assignment of the TPMs C. For instance, if $C = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$ for some permutation $\sigma : \mathcal{A} \to \mathcal{A}$, then $P^a_C = P_{\sigma(a)}$. Also, d_a and i_a denote the component corresponding to arm a in $\underline{d}$ and $\underline{i}$ respectively, and for $d \in \mathbb{N}$ and $i, j \in \mathcal{S}$, the notation $(P^a_C)^d(j|i)$ stands for the (i,j)th entry of the matrix $(P^a_C)^d$ obtained by multiplying P^a_C with itself d times. For a detailed derivation of (5), we refer the reader to [8, Section IV].

IV. Converse: Lower Bound

We now present the first main result of this paper: a lower bound for (3). Given two probability distributions μ and ν on $\mathcal{S}$, the *Kullback–Leibler (KL) divergence* (also called the *relative entropy*) between μ and ν is defined as

$$D_{\mathsf{KL}}(\mu\|\nu) \coloneqq \sum_{i\in\mathcal{S}} \mu(i) \log \frac{\mu(i)}{\nu(i)}, \quad (6)$$

where, by convention, $0 \log \frac{0}{0} = 0$.

Proposition 1. *Fix an assignment of the TPMs C. Then,*

$$\liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}^{\pi}_{C}[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{T^{\star}(C)}, \quad (7)$$

where $T^{\star}(C)$ is given by

$$T^{\star}(C) = \sup_{\nu} \min_{C'\in\mathsf{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d}, \underline{i}, a) k_{CC'}(\underline{d}, \underline{i}, a), \quad (8)$$

with $k_{CC'}(\underline{d}, \underline{i}, a) \coloneqq D_{\mathsf{KL}}((P^a_C)^{d_a}(\cdot|i_a)\|(P^a_{C'})^{d_a}(\cdot|i_a))$, where $P(\cdot|i)$ stands for the ith row of P. In (8)*, the supremum is over all probability measures ν on $\mathbb{S} \times \mathcal{A}$.*

To prove Proposition 1, we first derive an analogue of the ubiquitous change-of-measure result [4, Lemma 18] for the setting of restless arms. Given $\epsilon > 0$, we then lower bound $\inf_{\pi\in\Pi(\epsilon)} \mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))]$ by the relative entropy between two Bernoulli distributions with parameters ϵ and $1-\epsilon$. Next, we obtain an upper bound for $\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))]$ in terms of $\mathbb{E}^{\pi}_{C}[\tau(\pi)]$. This involves deriving an analogue of Wald's identity for the setting of restless Markov arms. Combining the upper and the lower bounds for $\mathbb{E}^{\pi}_{C}[Z^{\pi}_{CC'}(\tau(\pi))]$, we arrive at (7). The details are given in [8, Appendix A].

A. A Flow Constraint

The question of whether there exists ν attaining the supremum in (8) remains open. Also, because this supremum is over *all* probability distributions on $\mathbb{S} \times \mathcal{A}$, which is a large class of distributions, the lower bound in (7) may not be achievable. It seems necessary to introduce additional constraints on ν to render the lower bound achievable. Indeed, given $\delta > 0$, suppose ν_δ is a probability distribution on $\mathbb{S} \times \mathcal{A}$ such that

$$\min_{C'\in\mathsf{Alt}(C)} \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu_\delta(\underline{d}, \underline{i}, a)\ k_{CC'}(\underline{d}, \underline{i}, a) \geq T^{\star}(C) - \delta.$$

One way to achieve the quantity on the left hand side of the above equation is to ensure that for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, the long-term fraction of times the triplet $(\underline{d}, \underline{i}, a)$ is observed matches with $\nu_\delta(\underline{d}, \underline{i}, a)$. It seems difficult to accomplish this in the absence of additional structure on ν_δ.

In a related problem of *odd* arm identification, the authors of [6] are confronted with a similar difficulty in showing the achievability of the lower bound therein. To ameliorate the difficulty, they introduce a version of the following *flow* constraint on ν: for all $(\underline{d}', \underline{i}') \in \mathbb{S}$,

$$\sum_{a=1}^{K} \nu(\underline{d}', \underline{i}', a) = \sum_{(\underline{d},\underline{i})\in\mathbb{S}} \sum_{a=1}^{K} \nu(\underline{d}, \underline{i}, a)\ Q_C(\underline{d}', \underline{i}'|\underline{d}, \underline{i}, a). \quad (9)$$

In (9), $Q_C(\underline{d}', \underline{i}'|\underline{d}, \underline{i}, a)$ is the probability of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ making a transition from the state $(\underline{d}, \underline{i})$ to the state $(\underline{d}', \underline{i}')$ when arm a is selected; see [8] for an exact expression. With an end goal of showing achievability of the lower bound (7), we take cues from [6] and restrict attention to those probability distributions ν which satisfy (9).

In the next section, we design a policy that (a) is computationally feasible/tractable and easy-to-implement, and (b) achieves the lower bound in (7) when the supremum in (8) is evaluated over the restricted class of all ν satisfying (9).

V. Achievability: A Policy for Best Arm Identification

Notice that the outer summation in (8) is over all $(\underline{d}, \underline{i}) \in \mathbb{S}$. Because the arm delays are positive integer-valued, it follows that $\mathbb{S}$ is countably infinite and therefore cannot be stored on a machine with finite-size memory. As a result, any policy that operates on $\mathbb{S}$ is not practically implementable; see, for instance, the policy in [6]. To alleviate the difficulty arising from the countably infinite-valued arm delays, we study a simplified setting where the maximum delay of each arm is restricted to be at most R for some $R \in \mathbb{N} \cap (K, \infty)$,[1] and an arm whose delay is equal to R at any given time is forcibly selected in the following time instant. We write $\mathbb{S}_R$ to denote

[1] We consider $R > K$ to be consistent with our assumption that each of the arms is sampled once in the first K time instants.

the subset of $\mathbb{S}$ in which the delay of each arm is no more than R. For $a \in \mathcal{A}$, we let $\mathbb{S}_{R,a}$ denote the subset of $\mathbb{S}_R$ in which the delay of arm a is equal to R. Notice that $\mathbb{S}_{R,a}$ is a finite set for each a, and that $\mathbb{S}_{R,a} \cap \mathbb{S}_{R,a'} = \emptyset$ for all $a' \neq a$.

When the delay of each arm is constrained to be no more than R, we have the following additional constraint on ν:

$$\nu(\underline{d}, \underline{i}, a) = \sum_{a'=1}^{K} \nu(\underline{d}, \underline{i}, a') \quad \forall\, (\underline{d}, \underline{i}) \in \mathbb{S}_{R,a},\ a \in \mathcal{A}. \tag{10}$$

The constraint in (10) captures the fact that for any $a \in \mathcal{A}$, each occurrence of the state $(\underline{d}, \underline{i}) \in \mathbb{S}_{R,a}$ is followed by selecting arm a forcibly (i.e., with probability 1), thus implying that $\nu(\underline{d}, \underline{i}, a') = 0$ for all $a' \neq a$. Given an assignment of the TPMs C, let $T_R^\star(C)$ denote the value of

$$\sup_{\nu} \min_{C' \in \mathsf{Alt}(C)} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu(\underline{d}, \underline{i}, a)\ k_{CC'}(\underline{d}, \underline{i}, a), \tag{11}$$

where the supremum in (11) is over all probability distributions ν satisfying (10) and the flow constraint in (9) with $\mathbb{S}$ replaced by $\mathbb{S}_R$. Because the space of all probability measures on $\mathbb{S}_R \times \mathcal{A}$ is compact, and the objective function in (11) is continuous in ν, there exists $\nu^\star_{C,R}$ that attains the supremum in (11). Although a closed-form expression for $\nu^\star_{C,R}$ is not currently available, it can easily be evaluated numerically.

Before describing our policy for finding the best arm, we setup some notations. Fix $R \in \mathbb{N} \cap (K, \infty)$. Let π_R^{unif} denote the policy that selects the arms uniformly whenever the delay of each arm is $< R$, and forcibly selects an arm whose delay is equal to R. Under this policy, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a Markov process with $\mathbb{S}_R$ as its state space. The following result shows that this Markov process is, in fact, ergodic; the proof may be found in [8, Appendix 2].

Lemma 1. *Fix $R \in \mathbb{N} \cap (K, \infty)$. Under every assignment of the TPMs C, the process $\{(\underline{d}(t), \underline{i}(t)) :\ t \geq K\}$ is ergodic under the policy π_R^{unif}.*

Thanks to Lemma 1, the process $\{(\underline{d}(t), \underline{i}(t)) :\ t \geq K\}$ has a unique stationary distribution, say $\mu^{\mathsf{unif}}_{C,R}$, under the policy π_R^{unif} and under the assignment of the TPMs C. Clearly, $\mu^{\mathsf{unif}}_{C,R}(\underline{d}, \underline{i}) > 0$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}_R$. Let

$$\nu^{\mathsf{unif}}_{C,R}(\underline{d}, \underline{i}, a) \coloneqq \begin{cases} \frac{\mu^{\mathsf{unif}}_{C,R}(\underline{d}, \underline{i})}{K}, & (\underline{d}, \underline{i}) \notin \bigcup_{a'=1}^{K} \mathbb{S}_{R,a'}, \\ \mu^{\mathsf{unif}}_{C,R}(\underline{d}, \underline{i}), & (\underline{d}, \underline{i}) \in \mathbb{S}_{R,a}, \\ 0, & (\underline{d}, \underline{i}) \in \bigcup_{a' \neq a} \mathbb{S}_{R,a'}. \end{cases} \tag{12}$$

For $\eta \in (0, 1]$, let

$$\begin{aligned} \nu_{\eta,R,C}(\underline{d}, \underline{i}, a) &\coloneqq \eta\, \nu^{\mathsf{unif}}_{C,R}(\underline{d}, \underline{i}, a) + (1 - \eta)\, \nu^\star_{C,R}(\underline{d}, \underline{i}, a), \\ \mu_{\eta,R,C}(\underline{d}, \underline{i}) &\coloneqq \sum_{a=1}^{K} \nu_{\eta,R,C}(\underline{d}, \underline{i}, a). \end{aligned} \tag{13}$$

Clearly, $\mu_{\eta,R,C}(\underline{d}, \underline{i}) > 0$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}_R$. Let $\lambda_{\eta,R,C}$ be defined as

$$\lambda_{\eta,R,C}(a | \underline{d}, \underline{i}) \coloneqq \frac{\nu_{\eta,R,C}(\underline{d}, \underline{i}, a)}{\mu_{\eta,R,C}(\underline{d}, \underline{i})} \quad (\underline{d}, \underline{i}) \in \mathbb{S}_R,\ a \in \mathcal{A}. \tag{14}$$

Our policy, which we call *R-Delay-Constrained-Restless-BAI* (or R-DCR-BAI in short) or alternatively as $\pi^\star(L, \eta, R)$, is then as follows. Here, $L > 1$, $\eta \in (0, 1]$ and $R \in \mathbb{N} \cap (K, \infty)$ are parameters of the policy. The LLRs appearing in the policy description below are computed over the finite set $\mathbb{S}_R$.

Policy R-DCR-BAI **or** $\pi^\star(L, \eta, R)$:

Fix $L > 1$, $\eta \in (0, 1]$, and $R \in \mathbb{N} \cap (K, \infty)$. Assume that $A_0 = 1$, $A_1 = 2$, …, $A_{K-1} = K$. Let

$$M_C^{\pi^\star(L,\eta,R)}(n) \coloneqq \min_{C' \in \mathsf{Alt}(C)} Z_{CC'}^{\pi^\star(L,\eta,R)}(n).$$

Implement the following steps for all $n \geq K$.
(1) Compute $\bar{C}(n) \in \arg\max_{C \in \mathcal{C}}\ M_C^{\pi^\star(L,\eta,R)}(n)$. Resolve ties uniformly at random.
(2) If $M_{\bar{C}(n)}^{\pi^\star(L,\eta,R)}(n) \geq \log(L(K-1)(K-1)!)$, stop and declare the index of the best arm in $\bar{C}(n)$.
(3) If $M_{\bar{C}(n)}^{\pi^\star(L,\eta,R)}(n) < \log(L(K-1)(K-1)!)$, select arm A_n according to the distribution $\lambda_{\eta,R,\bar{C}(n)}(a | \underline{d}(n), \underline{i}(n))$.

In item (1) in R-DCR-BAI, $\bar{C}(n)$ denotes the estimate of the underlying assignment of the TPMs based on all the controls (arm selections) and observations seen up to time n. If the LLR between $\bar{C}(n)$ and its nearest alternative assignment of the TPMs exceeds a certain threshold (i.e., $\geq \log(L(K-1)(K-1)!)$), then the policy is sufficiently confident that $\bar{C}(n)$ is indeed the underlying assignment of the TPMs, and therefore stops and declares the index of the best arm in $\bar{C}(n)$. Else, it samples the next arm based on the value of $(\underline{d}(n), \underline{i}(n))$ according to $\lambda_{\eta,R,\bar{C}(n)}(\cdot | \underline{d}(n), \underline{i}(n))$.

VI. Results on the Performance of the Policy

We now present results on the performance of the policy R-DCR-BAI. Let $\pi^\star_{\mathsf{NS}}(L, \eta, R)$ denote a version of R-DCR-BAI that never stops, i.e., it does not check the second step in R-DCR-BAI and continues to the last step indefinitely.

Lemma 2. *Fix $L > 1$, $\eta \in (0, 1]$, and $R \in \mathbb{N} \cap (K, \infty)$. Given an assignment of the TPMs C, under $\pi = \pi^\star_{\mathsf{NS}}(L, \eta, R)$,*

$$\liminf_{n \to \infty} \frac{Z^\pi_{CC'}(n)}{n} > 0 \text{ almost surely},\ \forall\, C' \in \mathsf{Alt}(C). \tag{15}$$

Lemma 2 asserts that under $\pi = \pi^\star_{\mathsf{NS}}(L, \eta, R)$ and under the assignment of the TPMs C,

$$\liminf_{n \to \infty} \frac{M^\pi_C(n)}{n} > 0 \quad \text{almost surely}, \tag{16}$$

which implies that $M^\pi_C(n) \geq \log(L(K-1)(K-1)!)$ for all n large, almost surely. This proves that R-DCR-BAI stops in finite time with probability 1.

The next result shows that for an appropriate choice of the parameter L, the policy R-DCR-BAI achieves any desired error probability.

Lemma 3. *Fix an error probability threshold $\epsilon > 0$. If $L = 1/\epsilon$, then $\pi^\star(L, \eta, R) \in \Pi(\epsilon)$ for all $\eta \in (0, 1]$ and $R \in \mathbb{N} \cap (K, \infty)$. Here, $\Pi(\epsilon)$ is as defined in* (2)*.*

The below result establishes an almost sure upper bound on the stopping time of $\pi^\star(L,\eta,R)$.

Lemma 4. *Fix $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K,\infty)$. Given an assignment of the TPMs C, the stopping time of $\pi = \pi^\star(L,\eta,R)$ satisfies the almost sure upper bound*

$$\limsup_{L\to\infty} \frac{\tau(\pi^\star(L,\eta,R))}{\log L} \leq \frac{1}{\eta\, T_R^{\mathit{unif}}(C) + (1-\eta)\; T_R^\star(C)},$$

where

$$T_R^{\mathit{unif}}(C) \coloneqq \min_{C' \in \mathsf{Alt}(C)} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}_R} \sum_{a=1}^{K} \nu_{C,R}^{\mathit{unif}}(\underline{d},\underline{i},a)\; k_{CC'}(\underline{d},\underline{i},a).$$

The main result of this section, presented below, shows that the the almost sure upper bound of Lemma 4 also holds in expectation.

Proposition 2. *Fix $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K,\infty)$. Given an assignment of the TPMs C, the expected stopping time of the policy $\pi = \pi^\star(L,\eta,R)$ satisfies*

$$\limsup_{L\to\infty} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log L} \leq \frac{1}{\eta\, T_R^{\mathit{unif}}(C) + (1-\eta)\; T_R^\star(C)}. \tag{17}$$

Notice that the right hand side of (17) converges to $T_R^\star(C)$ as $\eta \downarrow 0$. Thus, the upper bound in (17) can be made arbitrarily close $T_R^\star(C)$ by choosing η suitably.

VII. A Key Monotonicity Property and Main Result

Before we present the main result of our paper, we note the following monotonicity property satisfied by $T_R^\star(C)$.

Lemma 5. *$T_R^\star(C) \leq T_{R+1}^\star(C)$ for all $R \in \mathbb{N} \cap (K,\infty)$.*

The proof of Lemma 5 is based on the observation that $\mathbb{S}_R \subset \mathbb{S}_{R+1}$ for all R. Noting that $T_R^\star(C) \leq T^\star(C)$ for all R, we get that $\lim_{R\to\infty} T_R^\star(C)$ exists.

With the above ingredients in place, we are now ready to state the main result of this paper.

Theorem 1. *Consider a multi-armed bandit with $K \geq 2$ arms in which each arm is a time homogeneous and ergodic discrete-time Markov process on the finite state space $\mathcal{S}$. Given TPMs $P_1,\ldots,P_K$ and a permutation $\sigma : \{1,\ldots,K\} \to \{1,\ldots,K\}$, let $C = (P_{\sigma(1)},\ldots,P_{\sigma(K)})$ be the underlying assignment of the TPMs where $P_{\sigma(a)}$ denotes the TPM of arm a. The growth rate of the expected time required to find the best arm in C satisfies the lower bound*

$$\liminf_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{\mathbb{E}_C^\pi[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{T^\star(C)}.$$

Further, given any $\epsilon > 0$, the policy $\pi^\star(1/\epsilon,\eta,R) \in \Pi(\epsilon)$ for all $\eta \in (0,1]$ and $R \in \mathbb{N} \cap (K,\infty)$. Additionally,

$$\limsup_{\substack{R\to\infty \\ \eta\downarrow 0 \\ L\to\infty}} \frac{\mathbb{E}_C^{\pi^\star(L,\eta,R)}[\tau(\pi^\star(L,\eta,R))]}{\log L} \leq \frac{1}{\lim_{R\to\infty}\; T_R^\star(C)}.$$

Thus, we see that the lower bound on the growth rate of the expected stopping time to find the best arm is $1/T^\star(C)$, and the upper bound is $1/(\lim_{R\to\infty}\; T_R^\star(C))$.

VIII. On the Equality of $\lim_{R\to\infty} T_R^\star(C)$ and $T^\star(C)$

Because $T_R^\star(C) \leq T^\star(C)$ for all $R \in \mathbb{N} \cap (K,\infty)$, it follows that $\lim_{R\to\infty} T_R^\star(C) \leq T^\star(C)$. Showing that, in general, this inequality is an equality appears to be difficult and remains open. One special case in which the inequality is indeed an equality is when the TPM of each arm has identical rows (i.e., each arm yields *i.i.d.* observations), all of which are identical to the stationary distribution associated with the TPM, as stated below. The proof is given in [8, Appendix I].

Lemma 6. *Suppose each row of P_k is equal to μ_k, $k = 1,\ldots,K$. In this special setting, $T^\star(C) = T_R^\star(C)$ for all $R \in \mathbb{N} \cap (K,\infty)$. Consequently, $\lim_{R\to\infty} T_R^\star(C) = T^\star(C)$.*

We are thus able to recover the results of [3], [4] for the case when the *i.i.d.* observations from each arm come from a finite alphabet.

IX. Discussion

- For any given $R \in \mathbb{N} \cap (K,\infty)$, the constant $T_R^\star(C)$ can be evaluated numerically. Also, the policy $\pi^\star(L,\eta,R)$ operates on the finite set $\mathbb{S}_R$ which can easily be stored in finite-size memory on a machine, thereby making it practically implementable.
- The function f in (1) used to define the best arm appears implicitly in the analyses of the lower and the upper bounds wherever one evaluates $\mathsf{Alt}(C)$ for any given C. By redefining $\mathsf{Alt}(C)$ appropriately, our analyses of the lower and the upper bounds may be extended to more general sequential hypothesis testing problems such as (a) finding the permutation σ such that the underlying assignment of the TPMs is $C = (P_{\sigma(1)},\ldots,P_{\sigma(K)})$ (here, $\mathsf{Alt}(C)$ is the set of all $C' \neq C$), (b) finding the second-best arm (here, assuming that the second-best arm is unique, $\mathsf{Alt}(C)$ is the set of all C' such that the index of the second-best arm in C' differs from that in C), etc.
- It will be interesting to extend our results to the more realistic setting in which decision entity only observes $Y_t^a = f(X_t^a)$ and not the underlying state X_t^a of arm a at time t, i.e., the arms yield *hidden Markov* observations. Here, f is the same function that appears in (1). Because $\{Y_t^a : t \geq 0\}$ is not a Markov process in general, the analyses of the lower and the upper bounds in this modified setting seem challenging and worth exploring.
- It will be interesting to extend the results of our paper to the case when the arm TPMs $P_1,\ldots,P_K$ are not known to the decision entity beforehand. Here, the arm TPMs must be estimated on-the-fly using the observations from the arms. In this case, showing that the TPM estimates converge to their true values is the key challenge.

Acknowledgements: This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018) and by the National Research Foundation Fellowship (A-0005077-01-00).

REFERENCES

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

[2] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *The Annals of Mathematical Statistics*, pp. 774–799, 1961.

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

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

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

[6] 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.

[7] ——, "Learning to detect an odd restless Markov arm with a trembling hand," *arXiv preprint arXiv:2105.03603*, 2021.

[8] P. N. Karthik, K. S. Reddy, and V. Y. F. Tan, "Best arm identification in restless Markov multi-armed bandits," 2022. [Online]. Available: https://arxiv.org/abs/2203.15236