Detecting an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik
Department of ECE,
Indian Institute of Science,
Bangalore - 560012

Rajesh Sundaresan
Department of ECE and
Robert Bosch Centre for Cyber Physical Systems,
Indian Institute of Science, Bangalore - 560012

***Abstract*—Consider a multi-armed bandit whose arms are independent Markov processes on a common underlying state space. The transition probability matrix of one of the arms (the odd arm) is different from the common transition probability matrix of all the other arms. The goal is to identify the odd arm as quickly as possible while keeping the probability of decision error small. We study the case of restless Markov observations and identify an asymptotic lower bound on the expected stopping time for a decision with vanishing error probability. We then propose a sequential test and show that the asymptotic behaviour of its expected stopping time comes arbitrarily close to that of the lower bound. Prior works dealt with iid arms and rested Markov arms, whereas our work deals with restless Markov arms.**

I. Introduction

The problem of odd arm identification deals with identifying an anomalous (or *odd*) arm in a multi-armed bandit as quickly as possible, while keeping the error probability small. Prior works on odd arm identification consider the cases when either each arm yields independent and identically distributed (iid) observations [1], [2], or when each arm yields (rested) Markov observations [3]. An important feature of the setting in [3] is that the Markov process of any given arm evolves by one time step only when the arm is selected, and does not evolve otherwise; this is known as the setting of *rested* arms. In this paper, we partially extend the results of [3] to the more difficult *restless* setting when the Markov process of each arm continues to evolve whether or not the arm is selected.

The continued evolution of the Markov process of each arm makes it necessary for the decision maker, whose objective it is to identify the index of the odd arm, to keep a record of (a) the time elapsed since each arm was previously selected (called the arm's *delay*), and (b) the state of each arm as observed at its previous selection time (called the *last observed state* of the arm). The notion of arm delays is absent when the arms are rested since the unobserved arms remain frozen at their previously observed states. Also, it is redundant when each arm yields iid observations since the last observed state of each arm is independent of the arm's current state. Therefore, the notion of arm delays introduces a new dimension to the problem, leads us to a certain countably infinite state space and the associated technical issues, and results in near optimal policies that potentially depend on history, unlike the cases in iid [1], [2] and rested Markov [3] settings.

A. Motivation and The Notion of a Trembling Hand

Our motivation to study the restless odd Markov arm problem comes from the desire to extend, to more general settings, the decision theoretic formulation of a certain visual search experiment conducted by Vaidhiyan et al. [1]. In this experiment, human subjects were shown a number of images at once, with one *oddball* image in a sea of *distracter* images. The goal of the experiment was to understand the relationship between (a) the time taken by the human subject to identify the oddball image, and (b) the dissimilarity between the oddball and distracter images as perceived by the human subject. The images used in the above experiment were static images. Vaidhiyan et al. also conducted experiments with dynamic drifting-dots images (movies), similar to the ones conducted by Krueger et al. [4], in which the dots in each movie location executed Brownian motions with identical drifts. Further, the drifts were identical in all the distracter movie locations, and were different from the drift in the oddball movie location. In this context, what are optimal strategies to identify the oddball movie? A systematic analysis of this question, along the lines of [1], requires an understanding of the restless odd Markov arm problem which is the main subject of this paper.

It is often the case in such visual search experiments that though the subject (or decision maker) has chosen a certain location to focus his attention, the actual focus location differs from the intended focus location with a small probability. We model this in our multi-armed bandit setting as a *trembling hand* for the decision maker: with probability $1 - \eta$, the decision maker pulls the intended arm, but with probability η, the decision maker pulls a uniformly randomly chosen arm.

In [5], we provide references to works that study the problem of *regret* minimisation in the context of restless Markov arms. In contrast, the problem considered in this paper is one of *optimal stopping* which, to the best of our knowledge, has not been studied elsewhere in the context of restless Markov arms. For more applications of the restless odd Markov arm problem, see [3]. For a related problem of best arm selection instead of odd arm selection, see [6], [7].

B. A Brief Overview of the Results

We show that the expected time to identify the odd arm, with probability of decision error at most ϵ, grows as $\Theta(\log(1/\epsilon))$. We identify the best constant multiplier, which we call

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

$R^*(P_1, P_2)$, in terms of the Markov transition probability matrices P_1 and P_2; the constant $R^*(P_1, P_2)$ is the best (smallest) growth rate. To do this, we first establish a lower bound on the growth rate. This uses the data processing inequality for the Kullback-Leibler divergence and a Wald-type lemma for Markov processes. Motivated by the structure of the optimisation in the lower bound, we study a family of Markov decision problems (MDPs). We then stitch together certain parameterised solutions to these MDPs and obtain a sequence of strategies whose growth rates come arbitrarily close to the lower bound.

In the concluding section, we discuss several insights.

II. Notations and Problem Formulation

We consider a multi-armed bandit with $K \geq 3$ arms, and define $\mathcal{A} \coloneqq \{1, \ldots, K\}$ to be the set of arms. We associate with each arm an ergodic, discrete time Markov process taking values from a finite state space $\mathcal{S}$. Further, we assume that the Markov process of each arm is independent of those of the other arms. The Markovian evolution of states on one of the arms (known as the *odd* arm) is governed by a transition probability matrix P_1, and that on each of the non-odd arms is governed by P_2, where $P_2 \neq P_1$. We denote by μ_i the unique stationary distribution of P_i, $i = 1, 2$.

For any integer $d \geq 1$ and a transition probability matrix P on $\mathcal{S}$, let P^d denote the transition probability matrix obtained by multiplying P with itself d times. For $i, j \in \mathcal{S}$ and $d \geq 1$, we write $P_1^d(j|i)$ and $P_2^d(j|i)$ to denote the (i, j)th element of the matrices P_1^d and P_2^d respectively (the case $d = 1$ corresponds to P_1 and P_2 respectively). We assume that for all $i, j \in \mathcal{S}$, (a) $P_1(j|i) > 0$ if and only if $P_2(j|i) > 0$, and (b) $\mu_1(i) > 0$ if and only if $\mu_2(i) > 0$. This assumption ensures that the decision maker cannot infer whether or not a given arm is the odd arm merely by observing certain specific state(s) on the arm. For $h \in \mathcal{A}$, we denote by $\mathcal{H}_h$ the hypothesis that h is the index of the odd arm.

We assume that P_1 and P_2 are known to a decision maker, whose goal it is to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. In order to do so, the decision maker devises a sequential arm selection strategy in which, at each discrete time instant $t \in \{0, 1, \ldots\}$, the decision maker first identifies an arm to pull; call this B_t (we use the phrases 'arm pulls' and 'arm selections' interchangeably). The decision maker however has a trembling hand and, as a consequence, the intended arm B_t gets pulled with probability $1 - \eta$ and a uniformly random arm gets pulled with probability η. The parameter η, which is fixed and strictly positive, governs the error in translating the decision maker's intention into an action. Write A_t for the arm that is pulled. The decision maker observes A_t, therefore knows whether or not his hand made an error in pulling the intended arm. Further, the decision maker observes the state of arm A_t, denoted by $\bar{X}_t$. The unobserved arms continue to undergo state evolution, making the arms *restless*. Thus, for each $t \geq 0$, B_t, A_t and $\bar{X}_t$ denote respectively the intended arm, the selected arm, and the observed state of the selected arm at time t. We use the shorthand notation $(B^t, A^t, \bar{X}^t)$ to denote the collection $(B_0, A_0, \bar{X}_0, \ldots, B_t, A_t\, \bar{X}_t)$.

A. Policy

A policy prescribes one of the following two actions at each time t: based on the history $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$,

- choose to pull the arm B_t according to a deterministic or a randomised rule, or
- stop and declare the index of the odd arm.

We use π to denote a generic policy, and let $\tau(\pi)$ denote the stopping time of policy π, where throughout this paper, all stopping times are defined with respect to the filtration $\mathcal{F}_t \coloneqq \sigma(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$, $t \geq 1$ and $\mathcal{F}_0 \coloneqq \{\Omega, \emptyset\}$. We write $\theta(\pi)$ to denote the index of the odd arm declared by the policy π at stoppage. Also, we write $P_h^\pi(\cdot)$ and $E_h^\pi[\cdot]$ to denote probabilities and expectations computed under policy π and under the hypothesis $\mathcal{H}_h$.

Given a target probability of error $\epsilon > 0$, we define $\Pi(\epsilon)$ as the set

$$\Pi(\epsilon) \coloneqq \{\pi : P_h^\pi(\theta(\pi) \neq h) \leq \epsilon \text{ for all } h \in \mathcal{A}\} \qquad (1)$$

of all policies whose probability of error at stoppage is below ϵ *for all possible odd arm locations* (since a policy does not know the true odd arm location, it has to work for all possible odd arm locations). We anticipate from similar results in the prior works that

$$\inf_{\pi \in \Pi(\epsilon)} E_h^\pi[\tau(\pi)] = \Theta(\log(1/\epsilon)).$$

Our interest is in characterising the constant factor multiplying $\log(1/\epsilon)$. For simplicity, we assume that for each $\epsilon > 0$, all policies in $\Pi(\epsilon)$ select each of the K arms in the first K time instants $t = 0, \ldots, K-1$, starting with arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K-1$. This does not affect the asymptotic analysis as $\epsilon \downarrow 0$.

B. 'Delays' in Observations

Recall that at each time $t = \{0, 1, \ldots\}$, the decision maker observes only one of the arms, while the unobserved arms continue to undergo state evolution. Therefore, the probability of the observing the state $\bar{X}_t$ on the *selected* arm A_t is a function of (a) the time elapsed since the previous time instant of selection of arm A_t (called the *delay* of arm A_t), and (b) the state of arm A_t at its previous selection time instant (called the *last observed state* of arm A_t). Notice that when the arms are *rested*, the notion of arm delays is absent since each arm remains frozen at its previously observed state until its next selection time instant. Also, the notion of arm delays is redundant in the setting of iid observations from the arms since, in this special case, the state of the selected arm at the current time is independent of its state at its previous selection. Thus, the notion of arm delays is a key distinguishing feature of the setting of restless arms.

We now define a new and more convenient notion of a state, based on the delays and the last observed states of the arms. As we show later, this new notion of state results in a Markov decision problem that is easier to comprehend.

For $t \geq K$, we denote by $d_a(t)$ and $i_a(t)$ respectively the delay and the last observed state of arm a at time t. Write $\underline{d}(t) := (d_1(t), \ldots, d_K(t))$ and $\underline{i}(t) := (i_1(t), \ldots, i_K(t))$ for the delays and the last observed states, respectively, of the arms at time t. Note that arm delays and last observed states are defined only for $t \geq K$ since these quantities are well-defined only when at least one observation is available from each arm. We set $\underline{d}(K) = (K, K-1, \ldots, 1)$, and follow the convention that $d_a(t) \geq 1$ for all $t \geq K$, and that $d_a(t) = 1$ if and only if arm a is selected at time $t-1$.

We follow the rule below for updating the arm delays and last observed states: if $A_t = a'$, then

$$d_a(t+1) = \begin{cases} d_a(t)+1, & a \neq a', \\ 1, & a = a', \end{cases}$$
$$i_a(t+1) = \begin{cases} i_a(t), & a \neq a', \\ \bar{X}_t, & a = a'. \end{cases} \tag{2}$$

where $\bar{X}_t$ is the state of the arm A_t at time t.

One thus has the sequence of intended arm pulls, actual arm pulls, observations, and states as follows. At each $t \geq K$, based on $\underline{d}(t), \underline{i}(t)$, choose to pull B_t; due to the trembling hand, observe that A_t is pulled; see its state $\bar{X}_t$; then form $\underline{d}(t+1), \underline{i}(t+1)$. This repeats until stoppage, at which time we have the declaration $\theta(\pi)$ (under policy π) as the candidate odd arm.

C. Problem Formulation

From the update rule in (2), it is clear that the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset $\mathbb{S}$ of the countable set $\mathbb{N}^K \times \mathcal{S}^K$, where $\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers. The subset $\mathbb{S}$ is formed by noting that at any time $t \geq K$, exactly one of the components of $\underline{d}(t)$ is equal to 1, and all the other components are strictly greater than 1. The evolution of the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is *controlled* by the sequence $\{B_t\}_{t\geq 0}$ of intended arm selections under policy π, thus making it a controlled Markov process with $\{B_t\}_{t\geq 0}$ as the sequence of controls; see [5] for a mathematical description of this controlled Markov process. Thus we are in a Markov decision problem (MDP) setting.

Let us now make precise the state space, action space, transition probabilities and our objective. The state space of the MDP is $\mathbb{S}$, with the state at time t denoted $(\underline{d}(t), \underline{i}(t))$. Note that A_{t-1} can be extracted from $(\underline{d}(t), \underline{i}(t))$. The action space of the MDP is $\mathcal{A}$, with action B_t at time t possibly depending on the previous actions B^{t-1} and the previous states $\{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\}$. (It is easy to see that this is equivalent to taking an action based on $(B^{t-1}, A^{t-1}, \bar{X}^{t-1})$.) The transition probabilities for the MDP are given by the trembling hand rule

$$P(A_t = a | B_t) = \frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{B_t = a\}}, \quad \forall a \in \mathcal{A}, \tag{3}$$

the law associated with the Markov arm A_t, and the update rule (2). In (3), $\mathbb{I}$ denotes the indicator function. In [5], we provide an exact formula for the transition probabilities.

Our objective, however, is nonstandard in the context of MDPs, and more in line with what information theorists study. We are interested in determining, for each hypothesis $\mathcal{H}_h$, the following:

$$\lim_{\epsilon \downarrow 0} \quad \inf_{\pi \in \Pi(\epsilon)} \quad \frac{E_h^\pi[\tau(\pi)]}{\log(1/\epsilon)}. \tag{4}$$

In the next section, we provide some preliminaries on MDPs. The terminologies used follow that of Borkar [8].

III. Preliminaries on MDPs

A policy π may be described completely by specifying $P^\pi(B_t \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\})$ for all $t \geq K$. We say that π is a *stationary randomised strategy* (SRS) if there exists a Cartesian product λ of the form

$$\lambda = \bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \lambda_{(\underline{d}, \underline{i})}, \tag{5}$$

with the component $\lambda_{(\underline{d}, \underline{i})}(\cdot)$ being a probability measure on $\mathcal{A}$, such that for all $t \geq K$ and $b \in \mathcal{A}$,

$$P^\pi(B_t = b \mid B^{t-1}, \{(\underline{d}(s), \underline{i}(s)), K \leq s \leq t\}) = \lambda_{(\underline{d}(t), \underline{i}(t))}(b).$$

Such an SRS π will be denoted π^λ. Note that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is indeed a Markov process under the SRS π^λ. Let Π_{SRS} denote the set of all SRS policies.

For convenience, we write $\lambda_{(\underline{d}, \underline{i})}(\cdot)$ as $\lambda(\cdot | (\underline{d}, \underline{i}))$ so that we may write λ itself in the more familiar form $\lambda(\cdot|\cdot)$.

The trembling hand parameter η will be strictly positive in this paper. An immediate consequence of having $\eta > 0$ is the following important property.

Lemma 1. *Let $\eta \in (0, 1]$. For every $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, the Markov process $\{\underline{d}(t), \underline{i}(t) : t \geq K\}$ under the policy π^λ is irreducible, aperiodic, positive recurrent, and hence ergodic.*

Proof: See [5, Appendix A] ■

Thus, under any $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, a unique stationary distribution exists for $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$; call it μ^λ.

IV. Lower Bound

We now present a lower bound for (4).

Given two probability distributions μ and ν on the finite state space $\mathcal{S}$, the Kullback-Leibler (KL) divergence between μ and ν is defined as

$$D(\mu \| \nu) := \sum_{i \in \mathcal{S}} \mu(i) \log \frac{\mu(i)}{\nu(i)}, \tag{6}$$

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

Proposition 1. *Fix $\eta \in (0, 1]$ and $h \in \mathcal{A}$. Assume that $\mathcal{H}_h$ is the true hypothesis. Let P_1 be the transition probability matrix of the Markov process of arm h, and for each $a \neq h$, let P_2 be the transition probability matrix of the Markov process arm a. Then,*

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h^\pi[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R^*(P_1, P_2)}, \tag{7}$$

where $R^*(P_1, P_2)$ is given by

$$R^*(P_1, P_2) := \sup_{\pi^\lambda \in \Pi_{\mathsf{SRS}}} \min_{h' \neq h} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\ k(\underline{d}, \underline{i}, a), \tag{8}$$

with

$$k(\underline{d}, \underline{i}, a) := \begin{cases} D(P_1^{d_a}(\cdot|i_a) \| P_2^{d_a}(\cdot|i_a)), & a = h, \\ D(P_2^{d_a}(\cdot|i_a) \| P_1^{d_a}(\cdot|i_a)), & a = h', \\ 0, & a \neq h, h'. \end{cases} \tag{9}$$

and

$$\nu^\lambda(\underline{d}, \underline{i}, a) := \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1 - \eta)\, \lambda(a|\underline{d}, \underline{i}) \right), \quad \forall (\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}. \tag{10}$$

Proof: See [5, Appendix B]. ■

The proof of the lower bound follows the outline in [3], with necessary modifications for the setting of restless arms. The key ingredients are the data processing inequality for relative entropies, a Wald-type lemma for Markov processes, and a recognition that the fraction of exits from the state $(\underline{d}, \underline{i})$ match the fraction of entries into that state. This forces the long-term probability of seeing the controlled Markov process in state $(\underline{d}, \underline{i})$ to be its unique invariant measure, by ergodicity (Lemma 1). These observations lead to an infinite dimensional convex programming problem with linear constraints, which is then transformed to (7).

Observe that the left hand side of (7) is evaluated by taking into consideration *all* policies, including those that are not necessarily SRS policies, whereas the supremum in (8) is only over SRS policies. This is a consequence of [9, Theorem 8.8.2]; see [5, Appendix B] for more details. Also, the right hand side of (7) is not a function of the odd arm index h. This is due to the symmetry in the structure of the arms.

V. Achievability

The question of whether the supremum in (8) is attained is still under study; see our remarks in the concluding section. Recall that this supremum is over all $\lambda(\cdot|\cdot)$ which are conditional probability distributions on the arms, conditioned on the arm delays and last observed states. This is in contrast with the works [1], [3], where the corresponding supremum is over all *unconditional* probability distributions on the arms. This is because, in those works, the arm delays are nonexistent. The unconditional probability measures are elements of the probability simplex on $\mathcal{A}$. The conditional probability measures are, however, more complex due to the countably many possible values for the arm delays. In spite of this added complexity, we can come arbitrarily close to the supremum in (8). We shall use this fact in our achievability result, which is the topic of this section.

We begin with some notations. Given $h, h' \in \mathcal{A}$, with $h \neq h'$, and a policy $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, let $Z_{hh'}(n)$ denote the log-likelihood ratio (LLR), under the policy π^λ, of all intended arm pulls, actual arm pulls, and observations up to time n under hypothesis $\mathcal{H}_h$ with respect to that under $\mathcal{H}_{h'}$. Writing P_h for $P_h^{\pi^\lambda}$, $Z_{hh'}(n)$ may be expressed as

$$\begin{aligned} Z_{hh'}(n) &= \log \frac{P_h(B^n, A^n, \bar{X}^n)}{P_{h'}(B^n, A^n, \bar{X}^n)} \\ &= \log \frac{P_h(B_0)}{P_{h'}(B_0)} + \log \frac{P_h(A_0|B_0)}{P_{h'}(A_0|B_0)} + \log \frac{P_h(\bar{X}_0|B_0, A_0)}{P_{h'}(\bar{X}_0|B_0, A_0)} \end{aligned} \tag{11}$$

$$+ \sum_{t=1}^{n} \log \left(\frac{P_h(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(B_t|B^{t-1}, A^{t-1}, \bar{X}^{t-1})} \right) \tag{12}$$

$$+ \sum_{t=1}^{n} \log \left(\frac{P_h(A_t|B^{t}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(A_t|B^{t}, A^{t-1}, \bar{X}^{t-1})} \right) \tag{13}$$

$$+ \sum_{t=1}^{n} \log \left(\frac{P_h(\bar{X}_t|A_t, B^{t}, A^{t-1}, \bar{X}^{t-1})}{P_{h'}(\bar{X}_t|A_t, B^{t}, A^{t-1}, \bar{X}^{t-1})} \right). \tag{14}$$

We now note that the probability of choosing arm B_t at time t, based on the history up to time t, cannot be a function of the underlying odd arm location (which is unknown), and must therefore be the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$. Thus, the first term in (11) and the expression in (12) are equal to 0. Also, for each t,

$$P_h(A_t|B_t, A^{t-1}, \bar{X}^{t-1}) = P_{h'}(A_t|B_t, A^{t-1}, \bar{X}^{t-1})$$

since A_t, the arm that is actually pulled at time t, is a function only of B_t and is related to B_t through (3). Therefore, given the history, the choice of A_t is not a function of the odd arm location, and is the same under hypotheses $\mathcal{H}_h$ and $\mathcal{H}_{h'}$, implying that the second term in (11) and the expression in (13) are 0. Finally, the probabilities in (14) do not depend on the intended arm pulls $\{B_t\}$ since the state $\bar{X}_t$ observed on arm A_t is a function only of the delay and the last observed state of arm A_t. Letting X_t^a denote the state of arm $A_t = a$, and defining

$$N(n, \underline{d}, \underline{i}, a) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a\}}, \tag{15}$$

$$N(n, \underline{d}, \underline{i}, a, j) := \sum_{t=K}^{n} \mathbb{I}_{\{\underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, A_t = a, X_t^a = j\}}, \tag{16}$$

for all $(\underline{d}, \underline{i}, a) \in \mathbb{S} \times \mathcal{A}$, $Z_{hh'}(n)$ may be expressed as

$$\begin{aligned} Z_{hh'}(n) &= \sum_{a=1}^{K} \log \frac{P_h(X_{a-1}^a)}{P_{h'}(X_{a-1}^a)} \\ &+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h, j) \log \frac{P_1^{d_h}(j|i_h)}{P_2^{d_h}(j|i_h)} \\ &+ \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{j \in \mathcal{S}} N(n, \underline{d}, \underline{i}, h', j) \log \frac{P_2^{d_{h'}}(j|i_{h'})}{P_1^{d_{h'}}(j|i_{h'})}. \end{aligned} \tag{17}$$

We now describe our policy. To do so, we first fix constants $\delta > 0$ and $L > 1$. These will be the parameters of our policy. Recall, from (8), that $R^*(P_1, P_2)$ is a supremum over all SRS policies. For a fixed hypothesis $\mathcal{H}_h$, by the definition of this supremum, we know that given $\delta > 0$, there exists $\lambda(\cdot \mid \cdot)$ such that under the corresponding SRS policy π^{λ}, we have

$$\min_{h' \neq h} \sum_{a=1}^{K} \sum_{(\underline{d},\underline{i}) \in \mathbb{S}} \nu^{\lambda}(\underline{d},\underline{i},a)\ k(\underline{d},\underline{i},a) > \frac{R^*(P_1,P_2)}{1+\delta}. \quad (18)$$

Notice that such a λ is, in general, a function of δ and the true hypothesis $\mathcal{H}_h$, although $R^*(P_1,P_2)$ itself is not a function of the true hypothesis; let us denote this λ as $\lambda_{h,\delta}$. Our policy, which we call $\pi^\star(L,\delta)$, is as below.

Policy $\pi^\star(L,\delta)$:
Fix $L > 1$ and $\delta > 0$. Let the trembling hand parameter be $\eta \in (0,1]$. Assume that $A_0 = 1$, $A_1 = 2$, and so on until $A_{K-1} = K$. Let $M_h(n) = \min_{h' \neq h} Z_{hh'}(n)$. Follow the below mentioned steps for each $n \geq K$.
(1) Pick an arbitrary $\theta(n) \in \arg\max_{h \in \mathcal{A}} M_h(n)$.
(2) If $M_{\theta(n)}(n) \geq \log((K-1)L)$, stop further arm selections and declare $\theta(n)$ as the true index of the odd arm.
(3) If $M_{\theta(n)}(n) < \log((K-1)L)$, decide to pull arm B_n according to the distribution $\lambda_{\theta(n),\delta}(\cdot \mid (\underline{d}(n), \underline{i}(n)))$.

In item (1) above, $\theta(n)$ denotes the guess of the odd arm at time n. In item (2), we check if the LLR of hypothesis $\mathcal{H}_{\theta(n)}$ with respect to each of its alternative hypotheses is separated sufficiently ($\geq \log(K-1)L$). If this is the case, then the policy is confident that the true odd arm index is $\theta(n)$. The policy then terminates and outputs the index $\theta(n)$. If the condition in item (2) fails, then the policy picks the next arm to observe.

Recall that the supremum in (8) is only over SRS policies. However, the policy $\pi^\star(L,\delta)$ described above is *not* an SRS policy since the distribution in item (3) is a function of $\theta(n)$ that could potentially depend on the entire history of arm selections and observations up to time n. Yet, as we show below, its performance comes arbitrarily close to that of the lower bound.

We now present results on the performance of our policy.

Lemma 2. *Fix $L > 1$, $\delta > 0$ and $h \in \mathcal{A}$, and suppose that $\mathcal{H}_h$ is the true hypothesis. Consider the non-stopping version of the policy $\pi^\star(L,\delta)$ which runs indefinitely (i.e., a policy that never stops and picks an arm at each time n according to the distribution in item (3) above). Under this policy, for every $h' \neq h$,*

$$\liminf_{n \to \infty} \frac{Z_{hh'}(n)}{n} > 0 \quad a.s.. \quad (19)$$

Proof: See [5, Appendix C]. ∎

As a consequence of Lemma 2, we have $\liminf_{n\to\infty} M_h(n)/n > 0$ under hypothesis $\mathcal{H}_h$ a.s.. This implies that $M_h(n) \geq \log((K-1)L)$ a.s. for all sufficiently large values of n, thus proving that the policy $\pi^\star(L,\delta)$ stops in finite time a.s..

Next, we show that the probability of error of our policy may be controlled by setting the parameter L suitably.

Lemma 3. *Fix error probability $\epsilon > 0$. If $L = 1/\epsilon$, then for every $\delta > 0$, $\pi^\star(L,\delta) \in \Pi(\epsilon)$.*

Proof: The proof uses the fact that the policy stops in finite time a.s.. See [5, Appendix D] for the details. ∎

With the above ingredients in place, we state the main result of this section, which is that the expected stopping time of our policy satisfies an asymptotic upper bound that comes arbitrarily close to the lower bound in (7).

Proposition 2. *Fix $h \in \mathcal{A}$ and $\delta > 0$, and let $\mathcal{H}_h$ be the true hypothesis. For $\pi = \pi^\star(L,\delta)$, the stopping time $\tau(\pi)$ satisfies*

$$\limsup_{L \to \infty} \frac{E_h^{\pi}[\tau(\pi)]}{\log L} \leq \frac{1+\delta}{R^*(P_1,P_2)}. \quad (20)$$

Proof: In the proof, which we provide in [5, Appendix E], we first show that as $L \to \infty$ (or equivalently $\epsilon \downarrow 0$), the ratio $\tau(\pi)/\log L$ satisfies an a.s. upper bound that matches with the right hand side of (20). We then show that the family $\{\tau(\pi)/\log L : L > 1\}$ is uniformly integrable. Combining the a.s. convergence with uniform integrability yields (20). ∎

VI. Concluding Remarks

We make several remarks to conclude the paper.

1. Since (20) holds for all $\delta > 0$, combining the assertions in Propositions 1 and 2, we see that

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h^{\pi}[\tau(\pi)]}{\log(1/\epsilon)} = \frac{1}{R^*(P_1,P_2)}. \quad (21)$$

Thus we have characterised the minimum growth rate of the expected stopping time, as $\epsilon \downarrow 0$ (see (4)).

2. The ergodicity of the Markov chain $(\underline{d}(t), \underline{i}(t))$ under any SRS policy π^{λ}, which is a consequence of having $\eta > 0$, is used in concluding that the time average approaches the ensemble average. This is crucial to show achievability. Note also the use of unique stationary distribution in the proof of the converse. The trembling hand model may be viewed as a *regularisation* that gives stability of the aforementioned Markov chain for free. If the trembling hand parameter η were 0, one could deliberately add some regularisation parameterised by η, and let this parameter $\eta \downarrow 0$. With $R^*_{\eta}(P_1,P_2)$ redefined as the growth rate with trembling hand parameter η, see (8), we note that $R^*_0(P_1,P_2)$ governs the lower bound, whereas $R^*_{0+}(P_1,P_2)$ governs the upper bound. This may result in a gap between the lower and upper bounds on the expected decision time. See [5, Section VII] for a more detailed discussion on the case $\eta = 0$.

3. The asymptotically optimal $\lambda(\cdot|\cdot)$ in the restless case may depend on history unlike the cases in [1], [3] where $\lambda(\cdot)$ did not depend on history, even in the rested Markov case. At first glance, this is surprising for the rested Markov case.

4. Computability of $R^*(P_1,P_2)$ may be an issue, and one must usually resort to Q-learning or other such simulation strategies to arrive at good policies. The fact that $D(P_i^{d_a}(\cdot|i_a)||P_j^{d_a}(\cdot|i_a))$ converges as $d_a \to \infty$ could enable restriction of the countable state space $\mathbb{S}$ to a finite set, and could lead to good approximations.

5. We have not studied achievability for case when $\eta = 0$. Naive extensions, based on the work in this paper, may lead to a gap between the upper and lower bounds as highlighted in the second point above. Another interesting case to study is the setting when P_1 and P_2 are unknown and have to be learnt along the way.

Acknowledgments

This work was supported in part by the Science and Engineering Research Board, Department of Science and Technology (grant no. EMR/2016/002503), by the Robert Bosch Centre for Cyber Physical Systems and the Centre for Networked Intelligence at the Indian Institute of Science.

References

[1] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 831–852, 2017.

[2] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Optimal odd arm identification with fixed confidence," *arXiv preprint arXiv:1712.03682*, 2017.

[3] P. N. Karthik and R. Sundaresan, "Learning to detect an odd markov arm," 2019. [Online]. Available: https://arxiv.org/abs/1904.11361

[4] P. M. Krueger, M. K. van Vugt, P. Simen, L. Nystrom, P. Holmes, and J. D. Cohen, "Evidence accumulation detected in bold signal using slow perceptual decision making," *Journal of neuroscience methods*, vol. 281, pp. 21–32, 2017.

[5] P. N. Karthik and R. Sundaresan, "Detecting an odd restless markov arm with a trembling hand (full version)," 2020. [Online]. Available: http://arxiv.org/abs/2005.06255

[6] 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.

[7] V. Moulos, "Optimal best markovian arm identification with fixed confidence," in *Advances in Neural Information Processing Systems*, 2019, pp. 5606–5615.

[8] V. S. Borkar, "Control of markov chains with long-run average cost criterion," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*. Springer, 1988, pp. 57–77.

[9] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.