Learning to Detect an Odd Markov Arm

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*—A multi-armed bandit with finitely many arms is studied when each arm is a homogeneous Markov process on an underlying finite state space. The transition law of one of the arms, referred to as the odd arm, is different from the common transition law of all other arms. A learner, who has no knowledge of the above transition laws, has to devise a sequential test to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. For this problem, we derive an asymptotic lower bound on the expected stopping time of any sequential test of the learner, where the asymptotics is as the probability of error vanishes. Furthermore, we 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 deal with iid arms, whereas our work deals with Markov arms.**

I. Introduction

We consider a multi-armed bandit with finitely many arms, in which each arm is identified with a homogeneous, irreducible and aperiodic discrete time Markov process, evolving on a common finite state space. The state evolution on one of the arms is according to a transition probability matrix P_1, while that on every other arm is according to P_2, where $P_1 \neq P_2$. The arm with transition matrix P_1 will be termed as the *odd arm*. A learner, who has no knowledge of P_1 or P_2, has to identify the index of the odd arm in the shortest possible time, while ensuring that the probability of identifying a wrong index is below a tolerance level $\epsilon > 0$.

The learner attempts to discover the unknown index of the odd arm by devising a sequential and adaptive arm selection scheme as follows: in every time slot, one out of the finitely many arms is selected; a state transition is observed on the selected arm; all other arms remain *rested* and do not exhibit state transitions in this time slot. The choice of which arm to select in any given time slot is based only on the history of arm selections and observations in all the previous time slots. For every such sequential strategy (or policy) of the learner, we aim to characterise the asymptotic behaviour of its expected stopping time, where the asymptotics is as $\epsilon \downarrow 0$. Identifying the optimal policy for a fixed $\epsilon > 0$ may be difficult [1, pp. 755]. As we shall see in this paper, the asymptotic analysis as $\epsilon \downarrow 0$ (see (8) and (22)) is tractable.

A. Prior Work

Our setting of rested and Markov arms is not very restrictive, and is shown in [2, Chapter 1] to closely model a host of real-life applications. The same setting also appears in Gittins's work [3] where it is assumed that the transition laws of each of the arms is known, and the goal is to devise policies that maximise the sum of average discounted rewards over an infinite duration of time. In a related problem of stochastic adaptive control, Agarwal et al. [4] strengthen Gittins's results to the case when the transition laws of the arms are not known, and are parametrised by an unknown parameter coming from a known, finite parameter space. While [3] and [4] deal with a problem of maximising reward or minimising regret over an infinite duration of time, ours is one of optimal stopping.

The problem of odd arm identification for iid observations can be embedded within the frameworks developed by Chernoff [1] and Albert [5]. There is a growing literature on this and related topics for iid observations. We provide a quick summary of only the most closely related works on odd arm identification. Vaidhiyan and Sundaresan [6] consider the special case of iid Poisson observations from each arm. Prabhu et al. [7] provide a more general treatment of iid observations coming from a generic exponential family. These works [5]–[7] provide lower bounds on the expected stopping time of any sequential policy for identifying the index of the odd arm, and in addition, also provide explicit schemes that achieve these lower bounds in the asymptotic regime as error probability vanishes. See [8]–[11] for many related works on iid observations. Our results are similar in spirit to those of [5]–[7], for the important setting of Markov arms.

B. Challenges in the Markov Setting

Markov observations offer some key challenges which must be overcome in the analysis. First, Wald's identity for iid settings, which greatly simplifies the analysis of the lower bounds in [6], [9], is not applicable. Next, we note that the scheme of [6] for iid Poisson observations is based on the important result [6, Prop. 3] that every arm may be chosen with a strictly positive probability. Such a result may not be available in other more general settings such as for Markov observations considered in this paper.

C. Contributions

We provide a lower bound for the expected stopping time of any policy that identifies the index of the odd arm without the knowledge of the transition matrices P_1 and P_2. This involves a generalisation of a result in [9] to the case of Markov observations since Wald's identity used in [9] is not applicable. We explicitly identify a configuration-dependent

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

constant in the lower bound that is a function of P_1 and P_2. This constant has the interpretation that it quantifies the effort required by any policy to learn the true index of the odd arm, by guarding itself against a nearest alternative with incorrect odd arm index. We then present a sequential scheme that is a modification of the generalised likelihood ratio test (GLRT). This modification is obtained by replacing the maximum appearing in the numerator of the usual GLR statistic with an average computed with respect to an artificial prior. We also borrow the idea of "forced exploration" from Albert's work [5] which guarantees that each arm is selected with a strictly positive probability, as in [6]. While this overcomes the need for proving the analogue of [6, Prop. 3] for Markov observations, it results in a penalty in the performance of our scheme. We show that this penalty can be made arbitrarily close to zero for a suitable choice of the forced exploration parameter. Using this, we show that the expected stopping time of our scheme can be made arbitrarily close to that given by the lower bound in the regime when $\epsilon \downarrow 0$. In [12], we provide simulation results showing the performance of our policy.

II. Notations

We fix $K \geq 3$, and consider a multi-armed bandit with K arms. We let $\mathcal{A} = \{1, 2, \ldots, K\}$ denote the set of arms. We associate with each arm a homogeneous, irreducible and aperiodic discrete time Markov process on a common finite state space $\mathcal{S}$, independent of the Markov processes of the other arms. Without loss of generality, let $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. Hereinafter, we use the phrase 'Markov process of arm a' to refer to the Markov process associated with arm $a \in \mathcal{A}$.

In every time slot $n = 0, 1, 2, \ldots$, one out of the K arms is selected and its state is observed. We let A_n denote the arm selected in slot n, and let $\bar{X}_n$ denote the state of arm A_n. We treat A_0 as the zeroth arm selection and $\bar{X}_0$ as the zeroth observation. Selection of an arm in slot n is based on the history $(A^{n-1}, \bar{X}^{n-1})$ of past observations and arms selected; here, $\bar{X}^k$ (resp. A^k) is a shorthand notation for the sequence $\bar{X}_0, \ldots, \bar{X}_k$ (resp. $A_0, \ldots, A_k$). We shall refer to such a sequence of arm selections and observations as a policy, which we generically denote by π. For each $a \in \mathcal{A}$, we denote the Markov process of arm a by $(X_k^a)_{k \geq 0}$. We denote by $N_a(n)$ the number of times arm a is selected by a policy up to (and including) slot n. Further, for each $a \in \mathcal{A}$ and states $i, j \in \mathcal{S}$, we denote by $N_a(n, i)$ and $N_a(n, i, j)$ respectively the number of times up to (and including) slot n the Markov process of arm a is observed to *exit* out of state i and to *exit* out of state i and *enter* into state j. Thus,

$$N_a(n) = \sum_{t=0}^{n} 1_{\{A_t = a\}}, \quad N_a(n, i) = \sum_{m=1}^{N_a(n)-1} 1_{\{X_{m-1}^a = i\}}, \tag{1}$$

$$N_a(n, i, j) = \sum_{m=1}^{N_a(n)-1} 1_{\{X_{m-1}^a = i,\, X_m^a = j\}}. \tag{2}$$

Our setting is one in which one of the arms is anomalous (hereinafter referred to as the odd arm). We let H_h denote the hypothesis that the index of the odd arm is $h \in \mathcal{A}$. We assume that the transition matrix of the Markov process of arm h is $P_1 = (P_1(j|i))_{i,j \in \mathcal{S}}$, while that of all other arms is $P_2 = (P_2(j|i))_{i,j \in \mathcal{S}}$; here, $P(j|i)$ is the entry in the ith row and jth column of matrix P. Further, we let μ_1 and μ_2 denote the unique stationary distribution of P_1 and P_2 respectively. We denote by ν the distribution of the initial state of each Markov process. In other words, for arm $a \in \mathcal{A}$, we have $X_0^a \sim \nu$, and this is the same distribution for all arms. We assume that the transition matrices and their associated stationary distributions are unknown to the learner.

We refer to the triplet $C = (h, P_1, P_2)$ as a configuration. For each $a \in \mathcal{A}$, we denote by $(Z_h^a(n))_{n \geq 0}$ the log-likelihood process of arm a under the above configuration. Using the notations introduced thus far, we have

$$Z_h^a(n) = \begin{cases} 0, & N_a(n) = 0, \\ \log \nu(X_0^a), & N_a(n) = 1, \\ \log \nu(X_0^a) & \\ \quad + \sum\limits_{m=1}^{N_a(n)-1} \log P_h^a(X_m^a | X_{m-1}^a), & N_a(n) \geq 2, \end{cases} \tag{3}$$

where $P_h^a(j|i)$ is the conditional probability under hypothesis H_h of observing state j on arm a given that state i was observed on arm a at its previous sampling instant. Clearly,

$$P_h^a(j|i) = \begin{cases} P_1(j|i), & a = h, \\ P_2(j|i), & a \neq h. \end{cases} \tag{4}$$

We denote by $(Z_h(n))_{n \geq 0}$ the log-likelihood process under hypothesis H_h of *all* observations and arm selections up to (and including) slot n. Then, using (3) and the independence of the Markov processes across arms, $Z_h(n) = \sum_{a=1}^{K} Z_h^a(n)$.

To be formal, the observation process $(\bar{X}_n)_{n \geq 0}$ and the arm selection process $(A_n)_{n \geq 0}$ are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, P)$. All stopping times are defined with respect to the filtration $(\mathcal{F}_n)_{n \geq 0}$ given by

$$\mathcal{F}_n = \sigma(A^n, \bar{X}^n), \quad n \geq 0. \tag{5}$$

Let $\tau(\pi) = \tau$ denote the stopping time of policy π. For each $a \in \mathcal{A}$ and each $i, j \in \mathcal{S}$, let $N_a(\tau)$, $N_a(\tau, i)$ and $N_a(\tau, i, j)$ denote the quantities in (1)-(2), with n replaced by τ.

We write $E^\pi[\cdot|C]$ and $P^\pi(\cdot|C)$ to denote expectations and probabilities computed under policy π, given that the true configuration of the arms is C. Given a tolerance parameter $\epsilon > 0$, our interest is in the class of policies whose probability of error at stoppage for any underlying configuration of the arms is at most ϵ. We denote this class of policies by $\Pi(\epsilon)$:

$$\Pi(\epsilon) = \left\{ \pi : P^\pi(I(\pi) \neq h | C) \leq \epsilon \ \forall \ C = (h, P_1, P_2) \right\}, \tag{6}$$

where $I(\pi)$ denotes the index of the odd arm output by policy π at stoppage. We re-emphasise that π cannot depend on the knowledge of P_1 or P_2, but could attempt to learn these along the way.

Remark 1. *Fix an odd arm index h, and consider the simpler case when P_1, P_2 are known, $P_1 \neq P_2$. Let $\Pi(\epsilon|P_1, P_2)$ denote the set of all policies whose probability of error at stoppage is within ϵ. From the definition of $\Pi(\epsilon)$ in* (6), *it follows that*

$$\Pi(\epsilon) = \bigcap_{P_1, P_2 : P_1 \neq P_2} \Pi(\epsilon|P_1, P_2). \tag{7}$$

That is, policies in $\Pi(\epsilon)$ work for any P_1, P_2, with $P_1 \neq P_2$. It is not a priori clear whether the set $\Pi(\epsilon)$ is nonempty. That it is nonempty for the case of iid observations was established in [1]. In this paper, we show that $\Pi(\epsilon)$ is nonempty even for the setting of rested and Markov arms.

III. Lower Bound

For any two transition probability matrices P and Q of dimension $|\mathcal{S}| \times |\mathcal{S}|$, and a probability distribution μ on $\mathcal{S}$, define $D(P||Q|\mu)$ as the quantity

$$D(P||Q|\mu) := \sum_{i \in \mathcal{S}} \mu(i) \sum_{j \in \mathcal{S}} P(j|i) \log \frac{P(j|i)}{Q(j|i)},$$

with the convention $0 \log 0 = 0$. The following proposition gives an asymptotic lower bound on the expected stopping time of any policy $\pi \in \Pi(\epsilon)$, as $\epsilon \downarrow 0$.

Proposition 1. *Let $C = (h, P_1, P_2)$ denote the true configuration of the arms. Then,*

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E^{\pi}[\tau(\pi)|C]}{\log \frac{1}{\epsilon}} \geq \frac{1}{D^*(h, P_1, P_2)}, \tag{8}$$

where $D^*(h, P_1, P_2)$ is a configuration-dependent constant that is a function only of P_1 and P_2, and is given by

$$\begin{aligned} &D^*(h, P_1, P_2) \\ &= \max_{0 \leq \lambda \leq 1} \left\{ \lambda \, D(P_1||P|\mu_1) + (1 - \lambda) \frac{(K-2)}{(K-1)} D(P_2||P|\mu_2) \right\}. \end{aligned} \tag{9}$$

In (9), *P is a transition probability matrix whose entry in the ith row and jth column is given by*

$$P(j|i) = \frac{\lambda \mu_1(i) P_1(j|i) + (1 - \lambda) \frac{(K-2)}{(K-1)} \mu_2(i) P_2(j|i)}{\lambda \mu_1(i) + (1 - \lambda) \frac{(K-2)}{(K-1)} \mu_2(i)}. \tag{10}$$

Proof: See Section VII-A of the online version [12]. ■

Our proof, while broadly following the outline of the proof of the lower bound in [9], requires a generalisation. Wald's identity for iid settings cannot be applied for Markov observations. Instead, a change of measure technique to generalise [9, Lemma 18] is used. Furthermore, for any arm $a \in \mathcal{A}$, the long run frequency of observing the arm exit out of state $i \in \mathcal{S}$ is equal to that of observing the arm enter into state i. We then note that this common frequency is the stationary probability of observing the arm in state i. This is the reason for the appearance, in (9), of the unique stationary distributions μ_1 and μ_2 of the odd arm and the non-odd arms respectively. This step is possible due to the rested nature of the arms, and may not hold in the general "restless" bandit setting in which the unobserved arms continue to undergo state transitions.

Remark 2. *The right hand side of* (9) *is a function only of the transition matrices P_1 and P_2, and does not depend on the index h of the odd arm. This is due to symmetry in the structure of arms. However, we retain the index h in $D^*(h, P_1, P_2)$ to be consistent with the notation $C = (h, P_1, P_2)$ used to denote arm configurations.*

Going further, we let λ^* denote the value of λ that achieves the maximum in (9). We then define $\lambda_{opt}(h, P_1, P_2) = (\lambda_{opt}(h, P_1, P_2)(a))_{a \in \mathcal{A}}$ as the probability distribution

$$\lambda_{opt}(h, P_1, P_2)(a) := \begin{cases} \lambda^*, & a = h, \\ \frac{1 - \lambda^*}{K - 1}, & a \neq h. \end{cases} \tag{11}$$

In the next section, we construct a policy that, at each time step, chooses arms with probabilities that match with those in (11) in the long run, in an attempt to reach the lower bound. While it is not a priori clear that this yields an asymptotically optimal policy, we show that this is indeed the case.

IV. Achievability

In this section, we present a scheme that is a modification of the classical generalised likelihood ratio (GLR) test. Suppose that each arm is selected once in the first K time slots. Note that this does not affect the asymptotic performance. Then, under configuration $C = (h, P_1, P_2)$, the log-likelihood process $Z_h(n)$ may be expressed for any $n \geq K$ as

$$\begin{aligned} Z_h(n) = &\sum_{a=1}^{K} \log \nu(X_0^a) + \sum_{i,j \in \mathcal{S}} N_h(n, i, j) \log P_1(j|i) \\ &+ \sum_{i,j \in \mathcal{S}} \sum_{a \neq h} N_a(n, i, j) \log P_2(j|i), \end{aligned} \tag{12}$$

from which the likelihood process under C, denoted by $f(A^n, \bar{X}^n|C)$, may be written as

$$\begin{aligned} f(A^n, \bar{X}^n|C) = &\prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j \in \mathcal{S}} (P_1(j|i))^{N_h(n,i,j)} \\ &\cdot \prod_{i,j \in \mathcal{S}} (P_2(j|i))^{\sum_{a \neq h} N_a(n,i,j)}. \end{aligned} \tag{13}$$

Let $\text{Dir}(1, \ldots, 1)$ denote a Dirichlet distribution with $|\mathcal{S}|$ parameters $\alpha_1, \ldots, \alpha_{|\mathcal{S}|}$ with $\alpha_j = 1$ for all $j \in \mathcal{S}$. Then, denoting by $\mathscr{P}(\mathcal{S})$ the space of all transition probability matrices of size $|\mathcal{S}| \times |\mathcal{S}|$, we specify a prior on $\mathscr{P}(\mathcal{S})$ using the above Dirichlet distribution as follows: for any $P = (P(j|i))_{i,j \in \mathcal{S}} \in \mathscr{P}(\mathcal{S})$, $P(\cdot|i)$ is chosen according to the above Dirichlet distribution, and is independent of $P(\cdot|j)$ for all $j \neq i$. Further, for any two matrices $P, Q \in \mathscr{P}(\mathcal{S})$, the rows of P are independent of those of Q. Then, it follows that under this prior, the joint density at (P_1, P_2) for $P_1, P_2 \in \mathscr{P}(\mathcal{S})$ is

$$\begin{aligned} \mathscr{D}(P_1, P_2) := &\prod_{i \in \mathcal{S}} \frac{\prod_{j \in \mathcal{S}} (P_1(j|i))^{\alpha_j - 1}}{B(1 \ldots, 1)} \prod_{i \in \mathcal{S}} \frac{\prod_{j \in \mathcal{S}} (P_2(j|i))^{\alpha_j - 1}}{B(1 \ldots, 1)} \\ &= \frac{1}{B(1, \ldots, 1)^{2|\mathcal{S}|}}, \end{aligned} \tag{14}$$

where $B(1, \ldots, 1)$ denotes the normalisation factor for the distribution $\mathrm{Dir}(1, \ldots, 1)$, and the second line above follows by substituting $\alpha_j = 1$, $j \in \mathcal{S}$.

We denote by $f(A^n, \bar{X}^n | H_h)$ the average of the likelihood in (13) computed with respect to the prior in (14). From the property that the Dirichlet distribution is the appropriate conjugate prior for the observation process,

$$f(A^n, \bar{X}^n | H_h) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i \in \mathcal{S}} \frac{B((N_h(n,i,j)+1)_{j \in \mathcal{S}})}{B(1, \ldots, 1)} \prod_{i \in \mathcal{S}} \frac{B((\sum\limits_{a \neq h} N_a(n,i,j)+1)_{j \in \mathcal{S}})}{B(1, \ldots, 1)}, \quad (15)$$

where $B((N_h(n,i,j)+1)_{j \in \mathcal{S}})$ above denotes the normalisation factor of a Dirichlet distribution with parameters $(N_h(n,i,j)+1)_{j \in \mathcal{S}}$. Let $\hat{P}_{h,1}^n$ and $\hat{P}_{h,2}^n$ denote the maximum likelihood estimates of transition matrices P_1 and P_2 respectively, under hypothesis H_h. Taking partial derivatives of the right hand side (13) with respect to $P_1(j|i)$ and $P_2(j|i)$ for each $i, j \in \mathcal{S}$, and setting the derivatives to zero, we get

$$\hat{P}_{h,1}^n(j|i) = \frac{N_h(n,i,j)}{N_h(n,i)}, \quad \hat{P}_{h,2}^n(j|i) = \frac{\sum\limits_{a \neq h} N_a(n,i,j)}{\sum\limits_{a \neq h} N_a(n,i)}. \quad (16)$$

Plugging these back into (13), we get the maximum likelihood of all observations and actions under hypothesis H_h:

$$\begin{aligned} \hat{f}(A^n, \bar{X}^n | H_h) &:= \max_{C=(h,\cdot,\cdot)} f(A^n, \bar{X}^n | C) \\ &= \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j \in \mathcal{S}} \left\{ \left(\frac{N_h(n,i,j)}{N_h(n,i)} \right)^{N_h(n,i,j)} \right. \\ &\qquad \left. \cdot \left(\frac{\sum\limits_{a \neq h} N_a(n,i,j)}{\sum\limits_{a \neq h} N_a(n,i)} \right)^{\sum\limits_{a \neq h} N_a(n,i,j)} \right\}. \end{aligned} \quad (17)$$

For any two hypotheses H_h and $H_{h'}$, where $h' \neq h$, we define the modified GLR statistic of hypothesis H_h with respect to hypothesis $H_{h'}$, along the lines of [6], as

$$M_{hh'}(n) := \log \frac{f(A^n, \bar{X}^n | H_h)}{\hat{f}(A^n, \bar{X}^n | H_{h'})}, \quad n = 0, 1, 2 \ldots \quad (18)$$

Thus, our modified GLR statistic is one in which the maximum in the numerator of the usual GLR statistic is replaced by an average computed over the space $\mathscr{P}(\mathcal{S})$ with respect to the artificial prior introduced in (14). Letting $M_h(n) := \min\limits_{h' \neq h} M_{hh'}(n)$ denote the modified GLR of hypothesis H_h with respect to its nearest alternative, we now describe our policy $\pi^\star(L, \delta)$. Here, L and δ are two parameters for the policy.

Policy $\pi^\star(L, \delta)$:
Fix $L \geq 1$ and $\delta \in (0,1)$. Let $(B_n)_{n \geq 1}$ be a sequence of iid Bernoulli(δ) random variables such that B_{n+1} is independent of the sequence $(A^n, \bar{X}^n)$ for all $n \in \{0, 1, 2, \ldots\}$. Choose each of the K arms once in the first K time slots $n = 0, \ldots, K-1$. For each $n \geq K-1$, at the end of slot n, follow the procedure described below.

(1) Let $h^*(n) = \arg\max\limits_{h \in \mathcal{A}} M_h(n)$, the index with the largest modified GLR after n slots; resolve ties uniformly at random.

(2) If $M_{h^*(n)}(n) < \log((K-1)L)$, choose the next arm A_{n+1} based on $(A^n, \bar{X}^n)$ as per the following rule:

(a) If $B_{n+1} = 1$, choose an arm uniformly at random.

(b) If $B_{n+1} = 0$, choose A_{n+1} according to the distribution $\lambda_{opt}(h^*(n), \hat{P}_{h^*(n),1}^n, \hat{P}_{h^*(n),2}^n)$, where for each $i, j \in \mathcal{S}$, the (i,j)th entries of the matrices $\hat{P}_{h^*(n),1}^n$ and $\hat{P}_{h^*(n),2}^n$ are as in (16), with h in (16) replaced by $h^*(n)$.

(3) If $M_{h^*(n)}(n) \geq \log((K-1)L)$, stop selections and declare $h^*(n)$ as the true index of the odd arm.

In the above policy, $h^*(n)$ is the best guess of the odd arm at the end of time slot n. If the modified GLR statistic of arm $h^*(n)$ is sufficiently larger than that of its nearest incorrect alternative ($\geq \log((K-1)L)$), then the learner is confident that $h^*(n)$ is the odd arm, stops taking further samples, and declares $h^*(n)$ as the odd arm. If not, the learner continues to obtain further samples.

We refer to the rule in item (2) above as *forced exploration* with parameter δ. A similar rule also appears in [5].

We now provide results on the performance of $\pi^\star(L, \delta)$. The main result on positive drift of the modified GLR statistic is as described in the following proposition.

Proposition 2. *Fix $L \geq 1$, $\delta \in (0,1)$, and consider the version of the policy $\pi^\star(L, \delta)$ that never stops. Let $C = (h, P_1, P_2)$ be the true configuration. Then, for all $h' \neq h$,*

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

Proof: The proof is based on the key idea that forced exploration with parameter $\delta \in (0,1)$ results in sampling each arm with a strictly positive rate that grows linearly with time. For details, see [12, Section VII-B]. ■

Prop. 2 forms the most important step in our analysis. It is in showing a similar result in [6] that the authors therein use their result of [6, Prop. 3] on guaranteed exploration at a positive rate. Indeed, it is not clear if this property holds in general. Instead, we appeal to [5] for the idea of forced exploration. See [8] on how to carry out forced exploration at a sublinear rate.

The result in (19) implies that for any given value of L, the modified GLR exceeds the threshold $\log((K-1)L)$ for some finite n, almost surely. Therefore, it follows from Prop. 2 that policy $\pi^\star(L, \delta)$ stops in finite time almost surely.

Next, we have the following result which shows that the parameter L may be set appropriately so that $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for any given choice of $\epsilon > 0$. We show later how δ may be chosen to achieve a desired performance of the policy.

Proposition 3. *Fix $\epsilon > 0$, $\delta \in (0,1)$. Then, for $L = 1/\epsilon$, we have $\pi^\star(L, \delta) \in \Pi(\epsilon)$.*

Proof: The proof uses Prop. 2 and the fact that policy $\pi^\star(L, \delta)$ stops in finite time almost surely. Further, the average in the numerator of the modified GLR statistic, in place of the maximum in the usual GLR statistic, is what enables us to show the result. For details, see [12, Section VII-C]. ■

The following proposition gives a more refined characterisation of the asymptotic drift of the process $(M_{hh'}(n)/n)_{n\geq 1}$.

Proposition 4. *Let $C = (h, P_1, P_2)$ denote the true configuration. Fix $\delta \in (0, 1)$. Then, under the non-stopping version of policy $\pi^\star(L, \delta)$, for any $h' \neq h$, we have*

$$\lim_{n\to\infty} \frac{M_{hh'}(n)}{n} = D^*_\delta(h, P_1, P_2) \quad a.s., \tag{20}$$

where the quantity $D^*_\delta(h, P_1, P_2)$ is given by

$$D^*_\delta(h, P_1, P_2) = \lambda^*_\delta \, D(P_1||P_\delta|\mu_1) + (1-\lambda^*_\delta)\frac{(K-2)}{(K-1)} D(P_2||P_\delta|\mu_2), \tag{21}$$

with $\lambda^*_\delta = \frac{\delta}{K} + (1-\delta)\lambda^*$, and for each $i, j \in \mathcal{S}$, $P_\delta(j|i)$ is as in (10) *with λ replaced by λ^*_δ.*

Proof: See [12, Section VII-D]. ■

Note that the policy $\pi^*(L, \delta)$ works with only estimated $\hat{P}^n_{h^*(n),1}$ and $\hat{P}^n_{h^*(n),2}$. To show (20), we must therefore ensure that the estimates approach the true values and a property akin to continuity holds, that is, taking actions based on $\hat{P}^n_{h^*(n),1}$ and $\hat{P}^n_{h^*(n),2}$, which are only approximately close to P_1 and P_2, adds only $o(1)$ to the drift $D^*_\delta(h, P_1, P_2)$.

With the above ingredients in place, we now have the following asymptotic upper bound on the performance of policy $\pi^\star(L, \delta)$.

Proposition 5. *Let $C = (h, P_1, P_2)$ denote the true configuration. Fix $\delta \in (0, 1)$. Then, under policy $\pi = \pi^\star(L, \delta)$,*

$$\limsup_{L\to\infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{1}{D^*_\delta(h, P_1, P_2)}. \tag{22}$$

Proof: The proof uses Prop. 4 and involves showing that the family $\{\tau(\pi^\star(L, \delta))/\log L : L \geq 1\}$ is uniformly integrable. For details, see [12, Section VII-E]. ■

Noting that $\lim_{\delta\downarrow 0} D^*_\delta(h, P_1, P_2) = D^*(h, P_1, P_2)$, it follows that $\delta \in (0, 1)$ may be chosen to ensure that the upper bound in (22) is as close to the lower bound in (8) as required.

V. Main Result

We now state the main result of this paper.

Theorem 1. *Let $C = (h, P_1, P_2)$ be the true configuration of the arms. Let $(\epsilon_n)_{n\geq 1}$ denote a sequence of error probability values satisfying $\epsilon_n \to 0$ as $n \to \infty$. Then, for each n and $\delta \in (0, 1)$, the policy $\pi^\star(L_n, \delta)$, with $L_n = 1/\epsilon_n$, belongs to the family $\Pi(\epsilon_n)$. Furthermore,*

$$\liminf_{n\to\infty} \inf_{\pi\in\Pi(\epsilon_n)} \frac{E[\tau(\pi)|C]}{\log L_n} = \lim_{\delta\downarrow 0} \lim_{n\to\infty} \frac{E[\tau(\pi^\star(L_n, \delta))|C]}{\log L_n} = \frac{1}{D^*(h, P_1, P_2)}. \tag{23}$$

Proof: The proof follows from the lower bound of Prop. 1 and from the results of Props. 3 and 5. ■

While those familiar with such stopping problems may easily guess the form of $D^*(h, P_1, P_2)$, the proof is not a straightforward extension of the iid case. To re-emphasise the challenges pointed out in Section I-B, Wald's identity is not available for the converse and a generalisation is needed, while a forced exploration approach provides achievability.

VI. Conclusions

We analyse the asymptotic behaviour of policies for a problem of odd arm identification in a multi-armed rested bandit setting with Markov arms. The asymptotics is in the regime of vanishing probability of error. Our setting is one in which the transition law of either the odd arm or the non-odd arms is not known. We derive an asymptotic lower bound on the expected stopping time of any policy as a function of error tolerance. We identify an explicit configuration-dependent constant in the lower bound. Furthermore, we propose a scheme that (a) is a modification of the GLRT, and (b) uses an idea of "forced exploration" from [5]. This scheme takes as inputs two parameters: $L \geq 1$ and $\delta \in (0, 1)$. We show that (a) for a suitable choice of L, the probability of error of our scheme can be controlled to any desired tolerance level, and (b) by tuning δ, the performance of our scheme can be made arbitrarily close to that given by the lower bound for vanishingly small error probabilities. In proving the above results, we highlight how to overcome some of the key challenges that the Markov setting offers in the analysis.

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

References

[1] H. Chernoff, "Sequential design of experiments," *Ann. Math. Stat.*, vol. 30, no. 3, pp. 755–770, 1959.

[2] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[3] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Royal Stat. Soc. Series B (Methodological)*, pp. 148–177, 1979.

[4] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space," *IEEE Trans. Autom. Control*, vol. 34, no. 12, pp. 1249–1259, 1989.

[5] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *Ann. Math. Stat.*, pp. 774–799, 1961.

[6] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 831–852, 2018.

[7] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an oddball target with observations from an exponential family," 2017. [Online]. Available: https://arxiv.org/abs/1712.03682

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

[9] 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.

[10] B. Hemo, K. Cohen, and Q. Zhao, "Asymptotically optimal search of unknown anomalies," in *Signal Processing and Information Technology (ISSPIT), 2016 IEEE Intl. Symposium on*. IEEE, 2016, pp. 75–80.

[11] S. Nitinawarat and V. V. Veeravalli, "Universal scheme for optimal search and stop," *Bernoulli*, vol. 23, no. 3, pp. 1759–1783, 08 2017.

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