Learning to Detect an Odd Restless Markov Arm

P. N. Karthik and Rajesh Sundaresan, Indian Institute of Science

***Abstract*— This paper studies the problem of identifying an anomalous arm in a multi-armed bandit when each arm is a finite-state Markov process and the arms are restless. Here, anomaly means that the transition probability matrix (TPM) of one of the arms (the odd arm) is different from the common TPM of each of the non-odd arms. The TPMs are unknown to a decision entity that wishes to find the index of the odd 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 odd arm index, where the asymptotics is as the error probability vanishes. Further, we devise a policy based on the principle of certainty equivalence, and demonstrate that under a continuous selection assumption and a regularity assumption on the TPMs, the policy achieves the lower bound asymptotically. Our achievability analysis is based on resolving the identifiability problem in the context of a certain countable-state controlled Markov process.**

I. Introduction

Consider a multi-armed bandit in which each arm is a time-homogeneous and ergodic Markov process evolving on a common, finite state space, and the arms are independent of each other. Suppose that the transition probability matrix (TPM) of one of the arms (the *odd* arm) is P_1, and that of each of the non-odd arms is P_2, where $P_2 \neq P_1$. A decision entity that knows neither P_1 nor P_2 wishes to find the index of the odd arm as quickly as possible, subject to an upper bound on the error probability (PAC formulation).

To arrive at a decision of the odd arm index, the decision entity samples the arms sequentially, one at each time $t \geq 0$. Following [1], we assume that the decision entity has a *trembling hand*. This means that for some fixed $\eta > 0$, the decision entity samples the intended arm with probability $1 - \eta$, but with probability η, the decision entity samples a uniformly randomly chosen arm; η is known as the trembling hand parameter. See [1] for an example where human subjects exhibit a trembling hand in a visual search experiment that involves searching for an oddball image in a sea of distracter images. It is likely that in such visual search experiments, the human subject scans multiple images at once before narrowing down the search to the oddball image [2], in a way different from that modelled in [1]. Nevertheless, the model in [1] captures the search dynamics in a way that makes the problem amenable to analysis. We therefore follow the model in [1].

At each time t, the decision entity observes the (noiseless) state of the pulled arm post-trembling. The Markov processes of the unobserved arms continue to evolve, thus making the arms *restless* [3]. The decision entity continues to sample the arms sequentially until it is sufficiently confident of its estimate of the odd arm index, at which time it stops further sampling and declares its estimate of the odd arm index.

A. Observation 'Delays' and a Markov Decision Problem

The restless nature of the arms makes it necessary for the decision entity to maintain a record of (a) the time elapsed since each arm was previously sampled (called the arm's *delay*), and (b) the state of each arm as observed at its previous sampling instant (called the arm's *last observed state*). The integer-valued arm delays introduce a countably infinite dimension to the problem. As demonstrated in [1, Section I.C], the delays and the last observed states of the arms together form a *controlled Markov process* and, in turn, lead to a Markov decision problem (MDP) whose state space is countably infinite and action space is the set of arms. Further, as seen from [1, Section I.C], the transition probabilities of the MDP are stationary across time and are functions of the odd arm index, P_1 and P_2. However, the objective is not to maximise rewards (or minimise regret), as is typical in MDPs, but to find the odd arm index quickly and accurately.

B. Certainty Equivalence and Identifiability

When neither P_1 nor P_2 is known beforehand, the transition probabilities of the MDP may be treated as being parameterised by a triplet of unknowns consisting of (a) the odd arm index, (b) the TPM P_1 of the odd arm, and (c) the common TPM P_2 of each of the non-odd arms. Call this triplet an *arms configuration*. Because the true arms configuration is not known beforehand, it must be learnt along the way.

A commonly used approach to learn the true parameter governing the transition probabilities of an MDP is *certainty equivalence*. The idea behind this approach is to (a) maintain an estimate of the parameter at each time t, and (b) take an action at time t supposing that the estimated value is indeed the true parameter value. The key challenge here is to show that the estimates converge to the true parameter as $t \to \infty$, i.e., the system is *identifiable*. Sufficient conditions that ensure identifiability under certainty equivalence have been proposed in the literature. For finite-state MDPs, Mandl [4] proposed a condition (known as *Mandl's identifiability condition*) on the MDP transition probabilities for identifiability. However, it is not clear if Mandl's condition is sufficient for identifiability in countable-state MDPs. In [5], the authors consider the same problem as Mandl's, but for countable-state MDPs when Mandl's identifiability condition is relaxed. The authors of [5] show that under some regularity assumptions on the MDP transition probabilities, certainty equivalence based on maximum likelihood (ML) estimation renders the system identifiable.

In this paper, we use certainty equivalence with ML estimation as in [5] to learn the true arms configuration. Due to the presence of arm delays in the likelihood function, closed-form expressions for the ML estimates of the TPMs are not

978-1-5386-8209-8/21/$31.00 ©2021 IEEE

available. Nevertheless, we show that the system is identifiable under a mild regularity assumption on the TPMs.

C. Specific Contributions

We derive an asymptotic lower bound on the expected time required to find the odd arm index subject to an upper bound on the error probability, where the asymptotics is as the error probability vanishes. Further, we devise a policy based on certainty equivalence and show that its expected time to find the odd arm, meeting the desired error probability constraint, satisfies an asymptotic upper bound that is arbitrarily close to the lower bound. Our analysis of the upper bound is based on two key assumptions that guarantee identifiability: (a) the existence of a continuous selection of stationary control laws that are near-optimal for the lower bound (continuous as a function of P_1 and P_2), similar to that appearing in [6], and (b) the regularity of the TPMs of the odd arm and the non-odd arm Markov processes. As we will see, the converse (lower bound) is for all arms configurations, whereas achievability (upper bound) is shown only for a subset of arms configurations that satisfy the regularity condition.

For an account of the prior works on odd arm identification, see the extended version [7]. For a related problem of best arm identification, see [8], [9]. See [7] for a comparison of our work with prior works on restless arms and anomaly detection.

II. Notations and Preliminaries

Consider a multi-armed bandit with $K \geq 3$ arms, and let $\mathcal{A} = \{1, \ldots, K\}$ denote the set of arms. For each $a \in \mathcal{A}$, consider a discrete-time Markov process $\{X_t^a : t \geq 0\}$ associated with arm a that is time-homogeneous, ergodic and takes values in a common finite set $\mathcal{S}$. Assume that the Markov process of each arm is independent of those of the other arms. Let the TPM of one of the arms (the *odd* arm) be P_1, and that of each non-odd arm be P_2, where $P_2 \neq P_1$. Write $C = (h, P_1, P_2)$ to denote an arms configuration in which h is the odd arm index, P_1 is the TPM of the odd arm h, and $P_2 \neq P_1$ is the TPM of each non-odd arm. Let $\mathcal{H}_h$ denote the composite hypothesis that h is the odd arm index.

A decision entity that knows neither P_1 nor P_2 wishes to find the odd arm index as quickly as possible while keeping the probability of its decision error small. Although the unknowns in the problem are (a) the odd arm index, (b) P_1, and (c) P_2, the objective of the decision entity is to only find the odd arm index accurately. The decision entity samples the arms sequentially, one at each time $t \geq 0$. Let B_t denote the arm that the decision entity intends to sample at time t. The decision entity has a trembling hand, and the arm A_t is instead sampled at time t, where A_t and B_t satisfy the probabilistic relation

$$P(A_t = a | B_t = b) = \frac{\eta}{K} + (1 - \eta)\, \mathbb{I}_{\{a=b\}} \tag{1}$$

for some $\eta > 0$. The decision entity observes A_t and therefore knows whether its hand trembled at time t. Further, the decision entity observes the (noiseless) state of the sampled arm A_t, which we denote by $\bar{X}_t$. While the decision entity observes the state of only one arm at each time instant, the Markov processes of the other arms continue to evolve (*restless* arms).

Define a *policy* π of the decision entity as a collection of functions $\{\pi_t : t \geq 0\}$ such that $\forall\ t \geq 0$, π_t takes as input the history $(B_0, A_0, \bar{X}_0, \ldots, B_t, A_t, \bar{X}_t)$ and outputs one of the following:

- sample arm B_{t+1} according to a deterministic or a randomised rule.
- stop sampling and declare the odd arm index.

Let $\tau(\pi)$ and $\theta(\tau(\pi))$ denote respectively the stopping time of policy π and the odd arm index output by policy π at stoppage. Because the decision entity is oblivious to the underlying arms configuration, any sequential arm sampling policy of the decision entity must meet the error probability constraint for all arms configurations. Given an error probability $\epsilon > 0$, let $\Pi(\epsilon)$ denote the set of all policies whose error probability at stoppage is $\leq \epsilon$ for all arms configurations, i.e.,

$$\Pi(\epsilon) := \Bigg\{ \pi : P^\pi\Big(\theta(\tau(\pi)) \neq h \;\Big|\; C = (h, P_1, P_2)\Big) \leq \epsilon \quad \forall\ C = (h, P_1, P_2),\ h \in \mathcal{A},\ P_2 \neq P_1 \Bigg\}. \tag{2}$$

In (2) and throughout the paper, $P^\pi(\cdot|C)$ denotes probabilities computed under the policy π and the arms configuration C. Similarly, $E^\pi[\cdot|C]$ will be used to denote expectations.

We anticipate from the prior works that for every arms configuration $C = (h, P_1, P_2)$,

$$\inf_{\pi \in \Pi(\epsilon)} E^\pi[\tau(\pi)|C] = \Theta(\log(1/\epsilon)). \tag{3}$$

The constant multiplying $\log(1/\epsilon)$ is, in general, a function of C. Our interest is in characterising the best (smallest) constant factor multiplying $\log(1/\epsilon)$ in the limit as $\epsilon \downarrow 0$. For simplicity, we assume that under every policy, $A_0 = 1$, $A_1 = 2$ and so on until $A_{K-1} = K$. If this is not the case, the arms may be sampled uniformly until the above sequence of arm samples is observed. Such an exercise of sampling the arms to first see the above sequence will only result in a finite delay (independent of ϵ) almost surely when $\eta > 0$, and does not affect the asymptotic analysis as $\epsilon \downarrow 0$.

A. Delays, Last Observed States and an MDP

Arm delays and last observed states are striking features of the setting of restless arms. See Section I for the definitions of these terms. Following the notations in [1], let $d_a(t)$ and $i_a(t)$ respectively denote the delay and the last observed state of arm $a \in \mathcal{A}$ at time t. Let $\underline{d}(t) = (d_a(t) : a \in \mathcal{A})$ and $\underline{i}(t) = (i_a(t) : a \in \mathcal{A})$ denote the vectors of arm delays and last observed states at time t. The arm delays and the last observed states make sense when each arm is sampled at least once, and shall therefore be defined for $t \geq K$ keeping in mind that $A_0 = 1, \ldots, A_{K-1} = K$ under every policy. For $t = K$, we set $\underline{d}(K) = (K, K-1, \ldots, 1)$. This means that with reference to time $t = K$, arm 1 was sampled K time instants earlier (i.e., at $t = 0$), arm 2 was sampled $K - 1$ time instants earlier (i.e., at $t = 1$) and so on. The rule for updating

$(\underline{d}(t), \underline{i}(t))$, based on the value of A_t, is straightforward and is given in [1, Eq. (2)]. From the update rule, it follows that $d_a(t) \geq 1$, with $d_a(t) = 1$ if and only if $A_{t-1} = a$. Therefore, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ takes values in a subset, say $\mathbb{S}$, of $\{1, 2, \ldots\}^K \times \mathcal{S}^K$. The set $\mathbb{S}$ is countably infinite and includes among many others the constraint that, for each $t \geq K$, exactly one component of the vector $\underline{d}(t)$ is equal to 1 and all other components are strictly greater than 1.

From [1, Section I.C], we know that $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a *controlled Markov process* with $(\underline{d}(t), \underline{i}(t))$ regarded as the state at time t and B_t regarded as the control at time t. That is, we are in the setting of an MDP whose state space is $\mathbb{S}$, action space is $\mathcal{A}$, and the transition probabilities under an arms configuration C are as given in [7, Eq. (4)]. The transition probabilities of the MDP are parameterised by the arms configuration. Our objective, however, is nonstandard in the context of MDPs and more in line with what information theorists study. Given an arms configuration $C = (h, P_1, P_2)$, we are interested in determining the following:

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

B. SRS Policies and State-Action Occupancy Measures

Call a policy π a *stationary randomised strategy (SRS)* if there exists a Cartesian product λ of the form[1]

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

with $\lambda(\cdot|\underline{d}, \underline{i})$ being a probability measure on $\mathcal{A}$ for all $(\underline{d}, \underline{i}) \in \mathbb{S}$, such that B_t is sampled according to $\lambda(\cdot|\underline{d}(t), \underline{i}(t))$ for all $t \geq K$. Let such an SRS policy be denoted more explicitly as π^λ, and let Π_{SRS} be the space of all SRS policies. Clearly, $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a Markov process under every SRS policy. Further, [1, Lemma 1] shows that this Markov process is ergodic when $\eta > 0$, and therefore possess a unique stationary distribution. Let $\mu^\lambda = \{\mu^\lambda(\underline{d}, \underline{i}) : (\underline{d}, \underline{i}) \in \mathbb{S}\}$ be the stationary distribution under π^λ. Also, for $(\underline{d}, \underline{i}) \in \mathbb{S}$ and $a \in \mathcal{A}$, let

$$\nu^\lambda(\underline{d}, \underline{i}, a) \coloneqq \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1 - \eta)\, \lambda(a|\underline{d}, \underline{i}) \right) \tag{6}$$

denote the *ergodic state-action occupancy measure* under π^λ.

III. Converse: Lower Bound

We now present a lower bound for (4). Given a TPM P on $\mathcal{S}$, an integer $d \geq 1$, and $i, j \in \mathcal{S}$, let $P^d(j|i)$ denote the (i, j)th entry of the matrix P^d.

Proposition 1. *Under the arms configuration $C = (h, P_1, P_2)$,*

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

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

$$R^*(h, P_1, P_2) \coloneqq$$

$$\sup_{\pi^\lambda \in \Pi_{SRS}} \inf_{\substack{C' = (h', P_1', P_2'): \\ h' \neq h, \\ P_1' \neq P_2'}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\; k_{CC'}(\underline{d}, \underline{i}, a), \tag{8}$$

with

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

The infimum in (8) *is over all alternative arms configurations $C' = (h', P_1', P_2')$ satisfying (a) $h' \neq h$, and (b) $P_1' \neq P_2'$. Also, $D(\cdot\|\cdot)$ in* (9) *is the relative entropy functional.*

Proof: See the extended version [7, Appendix A]. ∎

The key ingredients in the proof of the lower bound are (a) a data processing inequality for the setting of restless arms based on a change of measure argument presented in [8], (b) a Wald-type lemma for Markov processes, (c) a recognition of the fact that for any $(\underline{d}, \underline{i}) \in \mathbb{S}$, the long-term fraction of exits from $(\underline{d}, \underline{i})$ matches the long-term fraction of entries to $(\underline{d}, \underline{i})$, and (d) restriction of the supremum in (8) to the class of SRS policies, which is possible thanks to an analogue of [10, Theorem 8.8.2] for countable-state controlled Markov processes. A formal statement of this theorem as applicable to this paper may be found in [1, Appendix H].

The computability of $R^*(h, P_1, P_2)$ is an issue because the presence of ergodic state-action occupancy measures inside the summation in (8) does not allow for the simplification of the inner infimum. In contrast, such a simplification is possible in the setting of rested arms (see [11, pp. 4337-4338]) where the notion of arm delays is superfluous. Further, it is not clear if there exists an optimal SRS policy π^λ that attains the outer supremum in (8). However, for each $\delta > 0$, there exists $\lambda = \lambda_{h, P_1, P_2, \delta}(\cdot|\cdot) \in \bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$ such that

$$\inf_{\substack{C' = (h', P_1', P_2'): \\ h' \neq h,\, P_1' \neq P_2'}} \sum_{(\underline{d}, \underline{i}) \in \mathbb{S}} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a)\, k_{CC'}(\underline{d}, \underline{i}, a) \geq \frac{R^*(h, P_1, P_2)}{1 + \delta}. \tag{10}$$

Call $\lambda_{h, P_1, P_2, \delta}$ a *δ-optimal solution* for the arms configuration $C = (h, P_1, P_2)$. More generally, let $\lambda_{h, P, Q, \delta}$ denote a δ-optimal solution for $C = (h, P, Q)$. Notice that one or more λ may satisfy (10), thus implying that multiple δ-optimal solutions may exist for each arms configuration. In the next section, we show that under some regularity on the choice of δ-optimal solutions for the various possible arms configurations, a time-varying policy based on certainty equivalence and δ-optimal solutions achieves the lower bound asymptotically.

IV. Achievability

We begin this section by stating two key assumptions that form the basis for the results to be stated later.

[1] Writing $\mathcal{P}(\mathcal{A})$ to denote the space of all probability distributions on $\mathcal{A}$, it follows that (5) is an element of the product space $\bigotimes_{(\underline{d}, \underline{i}) \in \mathbb{S}} \mathcal{P}(\mathcal{A})$.

A. Two Key Assumptions

Given $\delta > 0$, (10) suggests that in order to approach the constant $R^*(h, P_1, P_2)$ in the lower bound to within a factor of $1/(1+\delta)$, the arms must eventually be sampled according to $\lambda_{h,P_1,P_2,\delta}$ or one of the δ-optimal solutions for $C = (h, P_1, P_2)$. Because the unknown underlying arms configuration may be any one among the uncountably infinite collection $\{C = (h, P, Q) : h \in \mathcal{A}, P \neq Q\}$ of all possible arms configurations, a feasible option is to eventually sample the arms according to $\lambda_{h,P,Q,\delta}$ which is "close" to $\lambda_{h,P_1,P_2,\delta}$ when (P, Q) close to (P_1, P_2). We show this works under some regularity conditions.

Assumption 1 (Continuous selection)**.** *For each $\delta > 0$, there exists a selection of δ-optimal solutions $\{\lambda_{h,P,Q,\delta} : h \in \mathcal{A}, P \neq Q\}$ such that for each $h \in \mathcal{A}$, the mapping $(P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous under (a) the topology arising from the Euclidean metric on the domain set, and (b) the product topology on the range set.*

The paper [6] considers a similar assumption as above, but for a more general sequential hypothesis testing problem in multi-armed bandits, but for independent observations (see [6, Assumption A]). Also, the analogue of Assumption 1 for the maximisers, instead of δ-optimal solutions, holds in the settings of the prior works [11]–[13] as a consequence of Berge's maximum theorem [14]. Going further, for each $\delta > 0$, fix a selection $\{\lambda_{h,P,Q,\delta} : h \in \mathcal{A}, P \neq Q\}$ of δ-optimal solutions satisfying Assumption 1.

Let $\mathscr{P}(S)$ denote the space of all TPMs on the finite set $\mathcal{S}$. For $\bar{\varepsilon}^* \in (0, 1)$, let

$$\mathscr{P}(\bar{\varepsilon}^*) \coloneqq \{P \in \mathscr{P}(\mathcal{S}) : \ P \text{ is ergodic}, \ \forall\ d \geq 1, \ i, j \in \mathcal{S}, \\ P^d(j|i) > 0 \implies P^d(j|i) \geq \bar{\varepsilon}^*\}. \quad (11)$$

Eq. (11) defines the class of all ergodic $P \in \mathscr{P}(\mathcal{S})$ such that each non-zero entry of P^d is lower bounded by $\bar{\varepsilon}^*$ uniformly in d. Clearly, every ergodic P belongs to $\mathscr{P}(\bar{\varepsilon}^*)$ for some P-dependent $\bar{\varepsilon}^*$. To see this, fix an arbitrary ergodic P, and let $\mu = (\mu(j) : j \in \mathcal{S})$ be the unique stationary distribution for P. From [15, Theorem 4.9], we have $P^d(j|i) \to \mu(j) > 0$ as $d \to \infty$ for all $i, j \in \mathcal{S}$. Let $\mu_{\min} = \min_j \mu(j)$. Then, there exists D such that $\forall\ d \geq D$, each non-zero entry of P^d is lower bounded by $\mu_{\min}/2$. Further, let

$$p_{\min} \coloneqq \min\{P^d(j|i) > 0 : i, j \in \mathcal{S}, \ d < D\}.$$

Then, we have $P \in \mathscr{P}(\bar{\varepsilon}^*)$, with $\bar{\varepsilon}^* = \min\{p_{\min}, \mu_{\min}/2\}$. Our assumption however requires this to hold uniformly across all possible pairs P, Q that can arise in our problem. Specifically,

Assumption 2. *P, Q belong to a subset of TPMs such that each row of P is mutually absolutely continuous with the corresponding row of Q. Additionally, there exists $\bar{\varepsilon}^* \in (0, 1)$ such that for all $C = (h, P, Q)$, the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$.*

Some remarks are in order. An arms configuration $C = (h, P, Q)$ satisfying Assumption 2 only increases the difficulty of identifying the odd arm index h. If $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ for some $\bar{\varepsilon}^* \in (0, 1)$, then they are harder to "distinguish" from one another. To see this, note that for any ergodic P, Q, we know from [15, Proposition 2.4] that that there exists $M = M(P, Q)$ such that all the entries of P^d and Q^d are strictly positive $\forall$ $d \geq M$, thus implying that $\forall\ d \geq M$ and $i \in \mathcal{S}$,

$$D(P^d(\cdot|i)\|Q^d(\cdot|i)) < \infty, \quad D(Q^d(\cdot|i)\|P^d(\cdot|i)) < \infty. \quad (12)$$

For $d < M$, it may be the case that one or both of the relative entropy terms in (12) equals $+\infty$ and discrimination becomes easier. However, when $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$, it follows that $\forall\ d \geq 1$, each row of P^d is mutually absolutely continuous with the corresponding row of Q^d. Furthermore, $\forall\ d \geq 1$ and $i, j \in \mathcal{S}$ such that $P^d(j|i) > 0$, $Q^d(j|i) > 0$, the relation

$$\bar{\varepsilon}^* \leq \frac{P^d(j|i)}{Q^d(j|i)} \leq \frac{1}{\bar{\varepsilon}^*} \quad (13)$$

holds, thus implying that $\forall d \geq 1$, each of the relative entropy terms in (12) is at most $\log(1/\bar{\varepsilon}^*)$. Therefore, $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ cannot have an arbitrarily large separation (in terms of relative entropy) and are harder to distinguish from one another.

B. Test Statistic

We now introduce a test statistic and use it later to construct a policy based on certainty equivalence. The test statistic is based on a modification of the usual generalised likelihood ratio (GLR) test statistic in which the numerator of the usual GLR test statistic is replaced with an average likelihood computed with respect to an artificial prior. The details are as follows. Let $\mathcal{P}(\mathcal{S})$ denote the space of all probability distributions on the set $\mathcal{S}$, and let $\text{Dir}(\alpha_j : j \in \mathcal{S})$ denote the Dirichlet prior on $\mathcal{P}(\mathcal{S})$ with parameters $(\alpha_j : j \in \mathcal{S})$. In particular, let $\text{Dir}(\mathbf{1})$ denote the Dirichlet distribution with $\alpha_j = 1\ \forall\ j \in \mathcal{S}$. Let D denote the prior on $\mathscr{P}(\mathcal{S})$ induced by $\text{Dir}(\mathbf{1})$ when each row of $P \in \mathscr{P}(\mathcal{S})$ is sampled independently according to $\text{Dir}(\mathbf{1})$.

Given $C = (h, P, Q)$, let $f(B^n, A^n, \bar{X}^n|C)$ denote the likelihood of all the arm samples and observations up to time n under the arms configuration C. Let $\bar{f}(B^n, A^n, \bar{X}^n|\mathcal{H}_h)$ denote the average likelihood of all the arm samples and observations up to time n under the hypothesis $\mathcal{H}_h$, where the averaging is over $(P, Q) \overset{iid}{\sim} D$. Further, let $\hat{f}(B^n, A^n, \bar{X}^n|\mathcal{H}_h)$ denote the maximum likelihood of all the arm samples and observations up to time n under the hypothesis $\mathcal{H}_h$, i.e.,

$$\hat{f}(B^n, A^n, \bar{X}^n|\mathcal{H}_h) = \sup_{P,Q} \ f(B^n, A^n, \bar{X}^n|C = (h, P, Q)). \quad (14)$$

Let $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ attain the supremum in (14). See [7] for remarks on (a) why the supremum in (14) is attained, and (b) why the constraint $P \neq Q$ is omitted in computing the supremum in (14). For $h, h' \in \mathcal{A}$ such that $h \neq h'$, our test statistic, which we denote by $M_{hh'}(n)$ at time n, is defined as

$$M_{hh'}(n) \coloneqq \log \frac{\bar{f}(B^n, A^n, \bar{X}^n|\mathcal{H}_h)}{\hat{f}(B^n, A^n, \bar{X}^n|\mathcal{H}_{h'})}. \quad (15)$$

The full expression for (15) is given in [7]. We refer to (15) as the *modified generalised likelihood ratio (GLR) test*

statistic of hypothesis $\mathcal{H}_h$ with respect to $\mathcal{H}_{h'}$. Let $M_h(n) = \min_{h' \neq h} M_{hh'}(n)$ denote the modified GLR test statistic of hypothesis $\mathcal{H}_h$ with respect to its nearest alternative hypothesis.

C. Policy

Fix $L > 1, \delta > 0$. Our policy, which we denote by $\pi^\star(L, \delta)$, is as below with L and δ as parameters.

Policy $\pi^\star(L, \delta)$:

Without loss of generality, let $A_0 = 1$, $A_1 = 2$, and so on until $A_{K-1} = K$. Follow the below mentioned steps $\forall\ n \geq K$.
(1) Compute $\theta(n) \in \arg\max_{h \in \mathcal{A}} \min_{h' \neq h} M_{hh'}(n)$. Resolve ties, if any, uniformly at random.
(2) If $M_{\theta(n)}(n) \geq \log((K-1)L)$, stop further sampling and declare $\theta(n)$ as the odd arm.
(3) If $M_{\theta(n)}(n) < \log((K-1)L)$, sample arm B_{n+1} according to $\lambda_{\theta(n), \hat{P}_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n), \delta}(\cdot \mid \underline{d}(n), \underline{i}(n))$.
(4) Update $n \leftarrow n+1$ and go back to (1).

In item (1) above, $\theta(n)$ denotes the policy's guess of the odd arm at time n. The certainty equivalence principle can be seen by noting that if the condition in item (2) fails, the policy supposes that $C_n = (\theta(n), \hat{P}^n_{\theta(n),1}(n), \hat{P}_{\theta(n),2}(n))$ is the true arms configuration, and samples arm B_{n+1} according to the δ-optimal solution for C_n. Observe that the policy does not rely on the knowledge of the constant $\bar{\varepsilon}^*$ from Assumption 2.

Remark 1. *The presence of arm delays in the expression for the maximum likelihood (see [7]) makes obtaining the closed-form expressions for $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ and the exact computation of $M_{hh'}(n)$ difficult. This difficulty can be overcome by repeatedly sampling each arm and estimating the TPMs using the consecutive observations obtained from the arms. However, given $\delta > 0$ and an arms configuration $C = (h, P_1, P_2)$, such an arm selection scheme may not lead to sampling the arms according to $\lambda_{h,P_1,P_2,\delta}$ eventually; such a sampling is crucial for obtaining an asymptotic upper bound that matches with the lower bound* (7)*. We show that this is indeed the case for the policy $\pi^\star(L, \delta)$ (Proposition 4) in spite of the computational intractability of $(\hat{P}_{h,1}(n), \hat{P}_{h,2}(n))$ and $M_{hh'}(n)$.*

D. Results on the Performance of the Policy

We now present the results on the performance of the policy $\pi^\star(L, \delta)$. See [7] for the proofs of these results. Throughout this section, we assume that Assumptions 1 and 2 hold.

Proposition 2. *Under the arms configuration $C = (h, P_1, P_2)$,*

$$\hat{P}_{h,1}(n) \longrightarrow P_1, \quad \hat{P}_{h,2}(n) \longrightarrow P_2 \quad as \quad n \to \infty \quad a.s..$$

The proof of Proposition 2 is based on verifying that the assumptions of [5] hold in the context of this paper. The result then simply follows from [5, Theorem 4.3].

Remark 2. *The proof of [5, Theorem 4.3] is based on a notion of "$\{\varepsilon_i\}$-randomisation" which, for controlled Markov processes, ensures a strictly positive probability of choosing each control at each time instant. The trembling hand model* (1) *of this paper ensures that the probability of sampling an arm at any given time is $\geq \frac{\eta}{K} > 0$, thus alleviating the need to consider $\{\varepsilon_i\}$-randomisations.*

Let $\pi^*_{\mathsf{ns}}(L, \delta)$ denote the non-stopping version of the policy, i.e., a policy that never stops and picks an arm at each time t according to the rule in item (3).

Proposition 3. *Fix $L > 1$ and $\delta > 0$. Let $C = (h, P_1, P_2)$ be the underlying arms configuration. Under the policy $\pi^\star_{ns}(L, \delta)$, $\forall\ h' \neq h$, we have*

$$\liminf_{n \to \infty} \frac{M_{hh'}(n)}{n} > 0 \quad a.s.. \tag{16}$$

An immediate consequence of Proposition 3 is that, a.s., $\liminf_{n\to\infty} M_h(n) > 0$, which in turn implies that, a.s., (a) the policy stops in finite time, and (b) for all t sufficiently large, $\theta(t) = h$ which, together with Proposition 2, establishes identification of the true arms configuration.

Let us now return to the policy $\pi^\star(L, \delta)$.

Proposition 4. *Fix an error probability $\epsilon > 0$. If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$.*

The main result of this section on the expected stopping time of the policy $\pi^\star(L, \delta)$ is as follows.

Proposition 5. *Fix $\delta > 0$. Under the arms configuration $C = (h, P_1, P_2)$, the expected stopping time of the policy $\pi = \pi^\star(L, \delta)$ satisfies*

$$\limsup_{L \to \infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}. \tag{17}$$

Consequently,

$$\limsup_{\delta \downarrow 0} \ \limsup_{L \to \infty} \frac{E^\pi[\tau(\pi)|C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}, \tag{18}$$

and $\pi = \pi^\star(L, \delta)$ is asymptotically optimal.

V. Concluding Remarks

This paper studied the problem of odd arm identification in restless multi-armed bandits when the transition probability matrices of the odd arm and the non-odd Markov processes are unknown. The main results of this paper are the following. (a) We gave an asymptotic lower bound on the expected time required to find the odd arm, subject to an upper bound on the error probability. The asymptotics is as the error probability vanishes. (b) We gave a policy based on the principle of certainty equivalence based on maximum likelihood (ML) estimates that achieves the lower bound asymptotically. The achievability analysis relies crucially on the convergence of the ML estimates of the transition probability matrices to their true values that is established under a continuous selection assumption and a certain regularity assumption.

Acknowledgments

This work was supported by the Science and Engineering Research Board, Department of Science and Technology (grant no. EMR/2016/002503).

REFERENCES

[1] P. N. Karthik and R. Sundaresan, "Detecting an odd restless markov arm with a trembling hand," *arXiv preprint arXiv:2005.06255*, 2020.

[2] M. Naghshvar and T. Javidi, "Two-dimensional visual search," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1262–1266.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

[4] P. Mandl, "Estimation and control in markov chains," *Advances in Applied Probability*, pp. 40–60, 1974.

[5] V. Borkar and P. Varaiya, "Identification and adaptive control of markov chains," *SIAM Journal on Control and Optimization*, vol. 20, no. 4, pp. 470–489, 1982.

[6] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality," *arXiv preprint arXiv:2007.12961*, 2020.

[7] P. N. Karthik and R. Sundaresan, "Learning to Detect an Odd Restless Markov Arm with a Trembling Hand: Full version," 2021. [Online]. Available: http://arxiv.org/abs/2105.03603

[8] 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.

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

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

[11] P. N. Karthik and R. Sundaresan, "Learning to detect an odd markov arm," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4324–4348, July 2020.

[12] 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.

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

[14] L. M. Ausubel and R. J. Deneckere, "A generalized theorem of the maximum," *Economic Theory*, vol. 3, no. 1, pp. 99–107, 1993.

[15] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.